

# Comparisons Between PISA and TIMSS – Are We the Man with Two Watches?

By  
Dougal Hutchison  
Ian Schagen  
National Foundation for Educational Research

*“A man with one watch always knows what time it is; a man with two watches is never quite sure” (Anon).*

## **Section 1. Introduction**

In 2002 and 2003 a strange thing started to happen in the UK. Government ministers and senior civil servants praised schools’ attainment - praise, moreover, based on the results of international studies. The Permanent Secretary at the DfES wrote:

*“For those doubters who constantly seek to run down (our education performance), we now have the OECD/PISA study – the biggest ever international study of comparative performance of 15-year-olds in 32 countries – which shows UK fourth in science, seventh in literacy and eighth in mathematics. Only Finland and Canada are consistently ahead of the UK – and major countries like Germany, Italy and Spain are well behind”.* (Normington, 2002).

Not all of those involved in the political process were equally impressed. Commenting on the PISA 2000 results, an Opposition MP stated:

*‘It is particularly incredible because in the previous year a far more authoritative study- the third international mathematics and science study, conducted by the respected International Association for the Evaluation of Educational Achievement,- put the UK 20<sup>th</sup> out of 41 countries...’* (Gibb, 2002).

So who is correct? Obviously the studies are not identical. But how different are they? In this presentation we describe the differences between TIMSS and PISA in a number of aspects. We also comment on the similarities which, at least numerically, seem at first sight to outweigh the differences. In this paper we will focus on TIMSS grade 8, and do not propose to cover the grade 4 studies.

## **Section 2: Frameworks of TIMSS and PISA**

### ***TIMSS***

TIMSS (Trends in Mathematics and Science Study) is the latest in a line of comparative international studies of attainment carried out under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). The IEA describes itself as an independent international co-operative of national research

institutions and governmental research agencies<sup>1</sup>. It is essentially a network of education researchers, though with a strong policy application. IEA has sponsored surveys in a number of topics, including reading comprehension, French, English, civics, ICT and literacy, but here we concentrate on mathematics and science. The first full study of mathematics took place in 1964, and the second sweep of studies of mathematics and science followed in the 1980s (Smithers, 2004). The Third International Mathematics and Science Study (TIMSS) was held in 1994-5, and a repeat was held in 1998-9. Further studies followed on a four year cycle in 2003, and one is being set up for 2007. The name TIMSS had stuck by this time, and rather than constantly renaming, the acronym was taken to stand for Trends in International Mathematics and Science Study. Forty-five countries took part in the 1995 TIMSS, 38 in 1999 and 49 in 2003 (Smithers, 2004, Mullis et al., 2004, 15). Costs in England are of the order of £1m (\$1.9m) (Bertrand, 2006).

Over this time IEA has gathered and perfected an enormous amount of knowledge and expertise in the organisation, design, administration and analysis of such studies. IEA and TIMSS are widely respected, rightly, for competence, integrity, innovation and relevance to the needs of the countries involved. To quote Brown (2000):

*'the three major international comparisons of mathematics attainment (carried out by) the IEA (...) have had a greater influence on education world-wide than any other single factor during the last 50 year period'.*

Later in this paper we shall discuss some of the technical aspects, such as sampling, test design and analysis, but at this stage it is worth paying tribute to the organisational and diplomatic skills involved in keeping on board so many different countries with their own customs, systems and agendas. In the earlier stages, anecdotal evidence suggested that stringency of standards could sometimes take second place to keeping countries on board, but at least in TIMSS, it is clear that the organisers have been able to insist on rather strict compliance with the rules of the studies.

It is difficult to find a single succinct statement of the aims of TIMSS. This is understandable: IEA studies have been around so long that they collect a range of aims, and cost so much it is obviously desirable that the subscribers can get as much as they can from it. There are statements of various degrees of prolixity in Robitaille et. al (1993) and TIMSS (2007). Robitaille et. al summarised the kinds of information sought by TIMSS (there is a comparable statement in Mullis *et al*, 2007):

- International variations in mathematics and science curricula, including variations in goals, intentions and sequences of curricula.
- International variations in the training of teachers in science and mathematics
- The influence of officially prescribed textbooks on the teaching of mathematics and science
- The course content that is actually taught in mathematics and science classrooms, i.e. opportunity to learn
- The effectiveness of different instructional practices

---

<sup>1</sup> International Association for the Evaluation of Educational Achievement Site Info Mission Statement [http://www.iea.nl/mission\\_statement.html](http://www.iea.nl/mission_statement.html) 21 September 2006

## Comparisons between PISA and TIMSS

- Students' achievement, especially in the area of non-routine problem solving and the application of science and mathematics in the 'real' world
- The attitudes and opinions of students and teachers
- The role of technology in the teaching and learning of science and mathematics, particularly the use of calculators and computers
- Participation rates in pre-university courses, with particular regard to gender-based differences
- The effect of tracking, streaming and other practices used to influence or direct students' course selection

Interestingly, actual comparison of attainment across countries is not explicitly mentioned above. At least half of the objectives carry an implication that the study aims to account for variation in learning. There is also a strong implication that students should be using mathematics and science as part of their role as a citizen, rather than as part of their employment armoury. Concentrating on measuring attainment, Mullis *et al* (2007) state that:

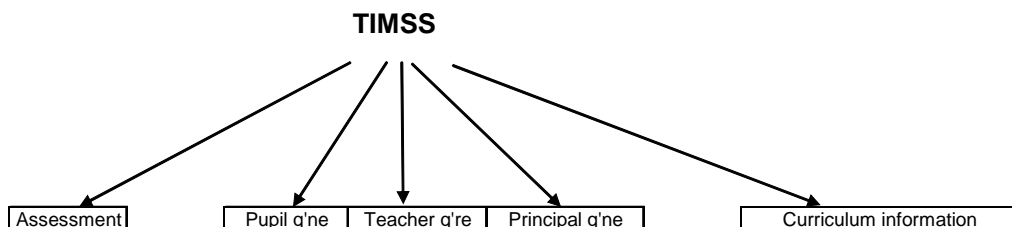
*'It is important (...) that students leaving high school are equipped with a fundamental understanding of science such that the decisions they make are informed ones. ... 'Prime reasons for having mathematics as a fundamental part of schooling include the increasing awareness that effectiveness as a citizen and success in a workplace are greatly enhanced by knowing and, more important, being able to use mathematics'.*

TIMSS is organised around two frameworks: a curriculum framework and an assessment framework. The curriculum framework<sup>2</sup> envisages three layers:

- Intended curriculum (what the teacher is expected to teach)
- Implemented curriculum (what the teacher taught)
- Attained curriculum (what the pupil learned).

Within the TIMSS assessment instruments (Figure 2.1), there are the two subject areas, mathematics and science, plus longer problem solving questions in mathematics and science from 2003, a student questionnaire, a teacher questionnaire and a questionnaire for school principals. Concentrating on the two subject areas, each is classified into a content dimension (domain) and a cognitive dimension (Figure 2.2).

FIGURE 2.1



<sup>2</sup> This has no counterpart in PISA.

Figure 2.2 TIMSS Assessment – Content & Cognitive Domains

Major Content Domain	Sub-domain	Cognitive Domains		
		Knowing	Applying	Reasoning
Mathematics	Number			
	Algebra			
	Geometry			
	Data & Chance			
Science	Biology			
	Chemistry			
	Physics			
	Earth Science			
(Problem solving)				

The mathematics **content** is classified along curriculum lines into Number, Algebra, Geometry and Data and Chance, while the science content is classified, again along syllabus lines, into Biology, Chemistry, Physics and Earth Science. The **cognitive** dimension is classified into Knowing, Applying and Reasoning. (Martin et. al, 2004)

### PISA

Compared with the long-established TIMSS series of studies, PISA, the Programme for International Student Assessment, is a relative newcomer. The PISA strategy was defined by participating countries (OECD, 1999), and has a governmental aura, since its studies are funded by OECD. PISA conducts its surveys on a three year cycle, with the first in 2000. It is not known precisely at present how much a sweep of PISA costs, but Canada estimated its participation as \$1.7m per year (CMEC, 2000). Costs in England are reported as comparable to those of TIMSS, but slightly higher (Bertrand, 2006). Rather than assessing the same subjects as TIMSS, PISA aims to assess reading literacy, mathematical literacy and scientific literacy, which it refers to as **domains**. As a mission statement, it is claimed that:

*‘The prime aim of the OECD/PISA assessment is to determine the extent to which young people have acquired the wider knowledge in reading literacy, mathematical literacy and scientific literacy that they will need in adult life.’* (OECD, 2004, 4).

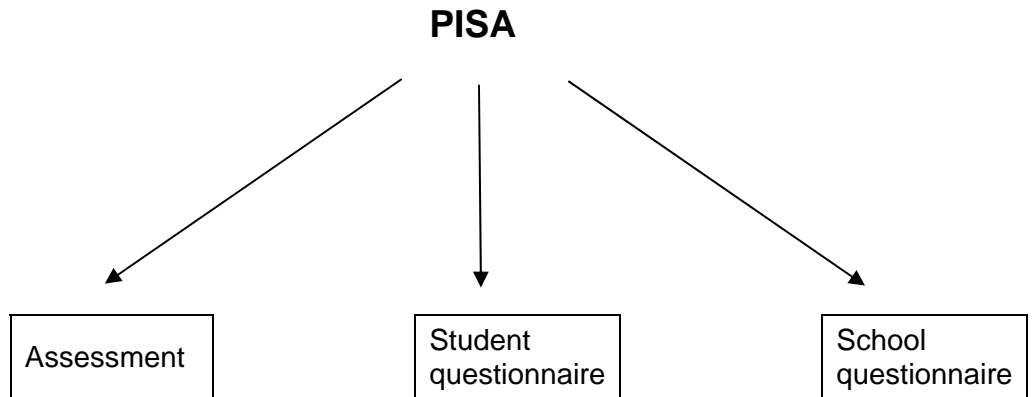
In each round, one domain is taken as the main subject, occupying about two thirds of the testing, with the remaining testing time being divided between the other two ‘minor’ domains. Thus in 2000, the main focus was Reading Literacy, in 2003 Mathematical Literacy, and 2006 Scientific Literacy. OECD also claims that PISA:

*‘provides insights into the factors that influence the development of the skills at home and at school and examines how these factors interact and what the implications are for policy development’* (OECD, 2003, 10).

Therefore, as with TIMSS, there is a suggestion of an attempt to find explanations for differences in performance.


Within the PISA framework, at the top level (Figure 2.3), there is the Pupil Assessment, a Student Questionnaire and a School Questionnaire for Principals. There is nothing corresponding to the TIMSS curriculum focus.

**FIGURE 2.3**



PISA assessments cover three domains, reading literacy, mathematical literacy and scientific literacy, plus problem solving<sup>3</sup> (Figure 2.4).

**Figure 2.4 PISA Assessment – Content & Process Domains**

Major Content Domain	Sub-domain	Process Domains		
		Knowing	Process	Context (Situation)
Mathematical Literacy				
Scientific Literacy	13 major themes (see below)			
Reading Literacy				
Problem solving				

Both mathematical and scientific literacy are classified in relation to **content, process and situations or contexts**. ‘Content’ and ‘process’ correspond substantially to TIMSS content and cognitive dimensions.

**Mathematical** literacy is divided into ‘big ideas’, Quantity, Shape and Space, Change and Uncertainty, and only secondarily into ‘curricular strands’ such as numbers, algebra and geometry. The definition of mathematical literacy is:

*“Mathematical literacy is an individual’s capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individuals’ life as a constructive, concerned and reflective citizen.”* (OECD, 2003, 24).

The other ‘big idea’ that comes across in the documentation is that of **mathematisation**, meaning converting a question from real life into a mathematical representation and solving and interpreting the resulting mathematical problem (OECD, 2003, 26).

**Scientific** literacy is defined as the capacity to use scientific knowledge to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity (OECD, 2003, 15). More extensively,

*“An important life skill (...) is the capacity to draw appropriate and guarded conclusions from evidence and information (...), to criticise claims made by others on the basis of the evidence put forward, and to distinguish opinion from evidence-based statements.”* (OECD, 1999, p. 59).

Within science literacy, the following major themes are identified (OECD, 2003, 136):

- Structure and properties of matter
- Atmospheric
- Chemical and physical change
- Energy transformations
- Forces and movement
- Form and function
- Human biology
- Physiological change
- Biodiversity
- Genetic control
- Ecosystems
- The Earth and its place in the Universe
- Geographical change .

The aims of TIMSS and PISA are explicitly different. Whereas TIMSS focused on the extent to which students have mastered mathematics and science as they appear in school curricula, PISA aimed to capture “the ability to use knowledge and skills to meet real-life challenges.” (OECD, 2001). In evidence to the House of Commons Education and Skills Committee, Barry McGaw (2002a), the Director for Education of the OECD, characterised the difference as TIMSS being interested to discover, “what science have you been taught and how much have you learned?”, while for PISA it was “what can you do with the science you have been taught?”

### 3. SAMPLING

#### *TIMSS – school sampling*

Smithers (2004) gives an interesting brief history of the IEA. In the early days, positive enhancement of the performance of the participating schools and pupils was, allegedly, not unknown. Where a country did particularly well (especially if this was unexpected) on the ground knowledge, or simply sour grapes, led to tales of irregularities in instruction, in sampling, and in administration of the test material, which materially improved the country's results. Strong political divisions precluded suggestions of *in situ* checking by unbiased outside personnel, and it is difficult to resist the conclusion that it was considered more important to keep everyone on side, and retain them in the sample, rather than risk alienating them by close investigation, even had this been possible. Each country carried out its own sampling.

The third round of IEA mathematics/science studies (TIMSS) in 1995 tightened this up substantially. For example, in 2003, the TIMSS sample was essentially specified by IEA, with central instructions, and specially provided software to carry out the sampling. Designs had to be carefully documented and agreed with the central coordinator before execution, and carefully documented while being drawn, with details being provided to the central site. All this, while doubtless damaging to the *amour propre* of countries which prided themselves on their sampling skills, knowledge and experience, was intended to provide a good, reliable sampling result from countries with a range of expertise and manpower at their disposal. The aim was to achieve a good combination of uniformity and sensitivity to local conditions. In this it has been largely successful, except that the more rigid sampling criteria (especially the insistence on a fixed minimum response rate) has disadvantaged countries (such as England) where there is good school-level data which can compensate for lower response rates. However, this is probably a debate for another time.

The country sample was essentially any country that wanted to take part and could afford the quite substantial costs. Other areas which could be defined as nations within countries (e.g. Scotland), or parts of countries (German Lander) also took part. The target population for the Grade 8 assessment was all students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing. This grade level was intended to represent eight years of schooling, counting from the first year of primary or elementary schooling, and was the eighth grade in most countries.

The international sample design for TIMSS is generally referred to as a two stage stratified cluster sample design. The first stage consists of a sample of schools, which may be stratified. Schools were sampled systematically from a comprehensive national list of all eligible schools with probabilities that were proportional to a measure of size - referred to as probability proportional to size (PPS) sampling<sup>4</sup>. The second stage consisted of a sample of one or more classrooms from the target grade in sampled schools. The sample size was defined on the basis of an effective sample

---

<sup>4</sup> PPS sampling produces a sample which is unbiased at the pupil level with respect to school size, when a fixed number of pupils are chosen per school.

## *Comparisons between PISA and TIMSS*

size of 400 pupils, after allowing for intracluster correlation. There was also a requirement to conduct analyses at school and classroom level, so at least 150 schools had to be selected. There was also a requirement for a minimum pupil sample size of 4000.

Schools could be excluded from the sampling frame for the following reasons:

- They were in geographically remote regions.
- They were of extremely small size.
- They offered a curriculum or a school structure that was different from the mainstream education system(s).
- They provided instruction only to students in the categories defined as “within-school exclusions”.

Certain types of students could be excluded, for example intellectually or functionally disabled students and those who had received less than one year of instruction in the language(s) of the test. Overall no more than 5 per cent of the total pupil population within the cohort could be excluded for the above reasons.

In order to be placed in the category of “acceptable sampling participation rate **without** the use of replacement schools”, a country had to have:

- a school response rate **without** replacement of at least 85% AND
- a student response rate of at least 85%.

An overall response rate of 75 % was required to be included in the report. There was provision for replacement schools, and some fairly complex arrangements for judging whether the eventual sample was acceptable.

### ***TIMSS – item sampling***

The programme aimed to assess a wide range of attainments. In order to cover all these, the pool of items and tasks included in the TIMSS assessment is extensive and would require much more testing time than could be allotted for individual students (about seven hours at grade 8). Therefore TIMSS 2003 and other years used a matrix-sampling technique that involves dividing the entire assessment pool into a set of unique item blocks, distributing these blocks across a set of booklets, and rotating the booklets among the students. Each student took one booklet containing both mathematics and science items. This design solicited relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses were aggregated across all respondents. This type of design had been used previously in the England and Wales Assessment of Performance Unit (APU) surveys (Foxman, Hutchison and Bloomfield, 1991) and in the US National Education of Educational Progress (NAEP) Surveys (Beaton, ed., 1988).

In the TIMSS 2003 assessment design, the 383 eighth-grade items were divided among 28 item blocks. The assessment time for individual students was 90 minutes (six 15-minute blocks). The booklets were organized into two three-block sessions (Parts I and II), with a break between the parts.

***PISA – school sampling***

The sample was restricted to OECD member countries, though in the event a number of ‘partner countries’ (e.g. Tunisia, Brazil) were also included. The desired PISA target population in each country consisted of 15-year-old students attending educational institutions located within the country, in grades 7 and higher. This included full-time and part time academic and vocational students. Home, on-the-job students and those not taking any type of education were excluded. This is unlikely to be a major problem in most developed OECD countries, but could well be relevant in the ‘partners’. In Mexico and Turkey, the official compulsory school age is 6-14. Even in such highly developed countries such as Germany and Switzerland, Prais (2003) argues that obligatory schooling is only up to age 15.

This meant that in all countries testing in April 2003, the national target population could have been defined as all students born in 1987 who were attending a school or other educational institution. A variation of up to one month in this age definition was permitted. If the testing was to take place at another time, the birth date definition had to be adjusted and approved by the consortium.

The sampling design used for the PISA assessment was a two-stage stratified sample in most countries. A minimum of 150 schools had to be selected in each country. The first-stage sampling units consisted of individual schools having 15-year-old students. In all but a few countries, schools were sampled systematically from a comprehensive national list of all eligible schools with probabilities proportional to size (PPS). In the selected schools 35 pupils were selected with equal probability from a list of all 15 year olds.

Schools could be excluded, for example, on the grounds of inaccessibility, or of removing a language group, possibly due to political, organisational or operational reasons. Certain types of students could be excluded, for example intellectually or functionally disabled students and those who had received less than one year of instruction in the language(s) of the test. A school attended only by students who would be excluded for intellectual, functional or linguistic reasons was considered as a school-level exclusion. It was required that the overall exclusion rate within a country be kept below 5 per cent.

A response rate of 85 per cent was required for initially selected schools. If the initial school response rate fell between 65 and 85 per cent, an acceptable school response rate could still be achieved through the use of replacement schools. To compensate for a sampled school that did not participate, where possible two replacement schools were identified for each sampled school. Furthermore, a school with a student participation rate between 25 and 50 per cent was not considered as a participating school for the purposes of calculating and documenting response rates. However, data from such schools was included in the database and contributed to the estimates included in the initial PISA international report.

A response rate of 80 per cent of selected students in participating schools was required. A student who had participated in the original or follow-up cognitive sessions was considered to be a participant. A student response rate of 50 per cent within each school was required for a school to be regarded as participating: the

overall student response rate was computed using only students from schools with at least a 50 per cent response rate.

Stratification varied between countries - schools were stratified geographically, or by such factors as type, size or academic attainment. Some countries requested some form of overlap control to avoid overlaps between TIMSS and PISA. Sample design was a long and complex procedure, with the defined sampled having to be agreed in detail with the central organisers. In general the central body actually drew the sample.

### ***PISA – item sampling***

As in TIMSS, a wide range of items was covered, though without any single student being asked to do them all, by the use of matrix sampling. The 167 main 2003 study items were allocated to 13 item clusters (seven mathematics clusters and two clusters in each of the other domains), with each cluster representing 30 minutes of test time. The items were presented to students in 13 test booklets, with each booklet being composed of four clusters. Each cluster appears in each of the four possible positions within a booklet exactly once.

### **Section 4. Test items**

Since the PISA 2006 study, the first with science as its major domain, is still ongoing, we concentrate on the 2003 study and on mathematics. For both TIMSS and PISA studies, the process of item development was long, complex and thorough, involving a number of different stages, including the field trials at which a larger number of items were tested before final item selection.

For PISA 2003, full details of this process can be found in the Technical Report (OECD, 2004, pp.15-32). The main study contained 85 mathematics items, categorised in terms of ‘competency clusters’ as follows:

- ‘Reproduction’ 26 items
- ‘Connections’ 40 items
- ‘Reflection’ 19 items

Split between the four major content categories, the numbers of items were:

- Space and shape 20 items
- Quantity 23 items
- Change and relationships 22 items
- Uncertainty 20 items

Finally, the numbers in the different item types were:

- Multiple choice 28 items
- Closed constructed response 13 items
- Open constructed response 44 items

### *Comparisons between PISA and TIMSS*

For TIMSS 2003, details of item development are to be found in Neidorf & Garden (2004). The main study contained 194 mathematics items, of which 115 were new and 79 were ‘trend’ items from previous surveys. The classification of items is similar to but subtly different from those for PISA; in terms of ‘mathematics cognitive domains’ the breakdown of items in 2003 was:

- Knowing facts and procedures      15%    (29 items)
- Using concepts                        20%    (39 items)
- Solving routine problems          40%    (78 items)
- Reasoning                                25%    (48 items)

In terms of the five mathematics content domains in TIMSS, the split was:

- Number                                    30%    (58 items)
- Algebra                                    25%    (49 items)
- Measurement                            15%    (29 items)
- Geometry                                 15%    (29 items)
- Data                                        15%    (29 items)

TIMSS only classifies items into two types; for TIMSS 2003 the numbers were:

- Multiple choice                        66%    (128 items)
- Constructed response                 34%    (66 items)

Clearly there are problems in making direct comparisons here between PISA and TIMSS, as the categories defined for the two studies are not the same. However, we have attempted to force such comparisons by conflating certain categories and trying as far as possible to compare like with like. The results of this procrustean procedure are illustrated in Figures 4.1 to 4.3, for the three broad classifications into cognitive/competency domains, content domains and item types.

We have combined the ‘knowing facts and procedures’ and ‘solving routine problems’ categories in TIMSS and put this alongside PISA’s ‘reproduction’ category – the graph indicates that this combined category tends to dominate the TIMSS items while the PISA items are more heavily weighted towards ‘connections’ (assumed parallel to ‘using concepts’ in TIMSS). ‘Reflection’ and ‘reasoning’ seem equally weighted in both studies.

**Figure 4.1: PISA and TIMSS 2003 Mathematics Cognitive/Competency Domains**

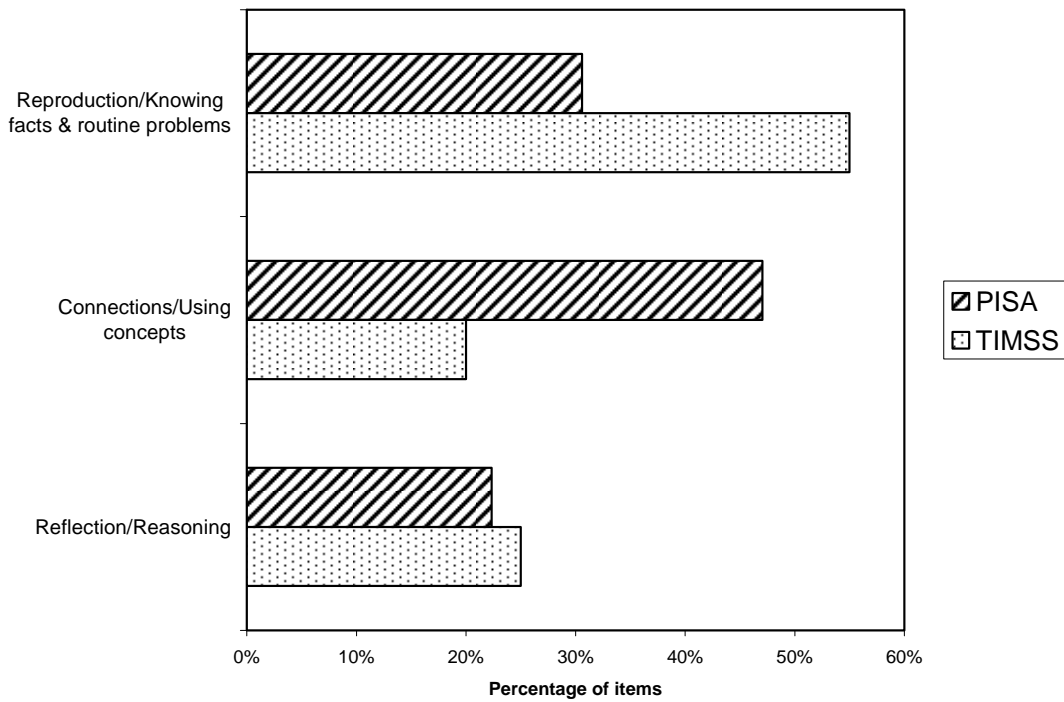
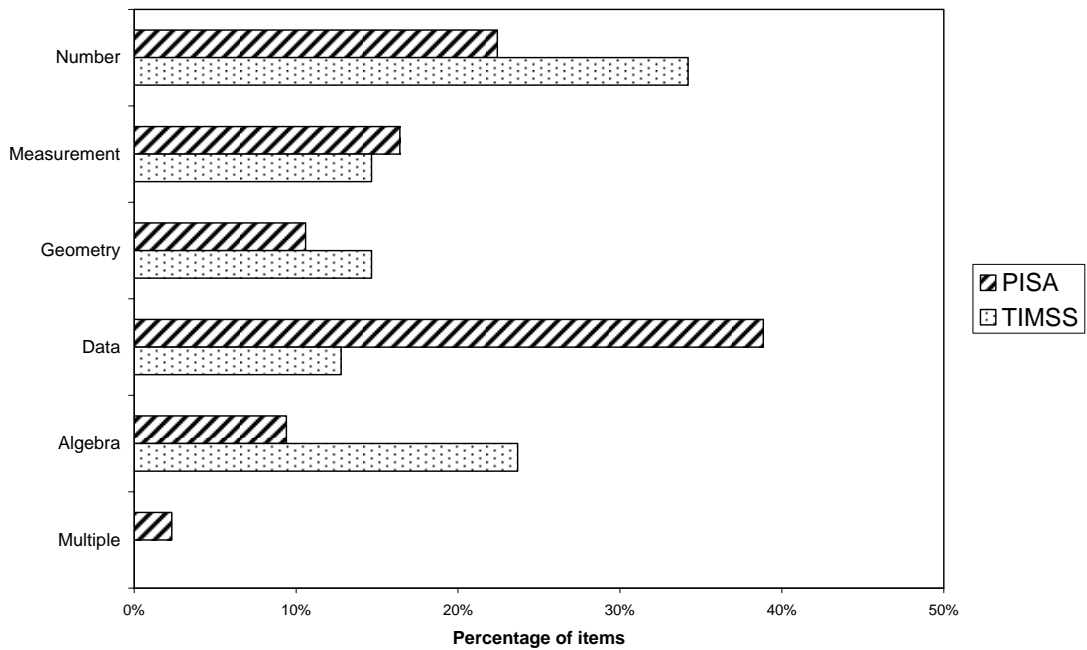


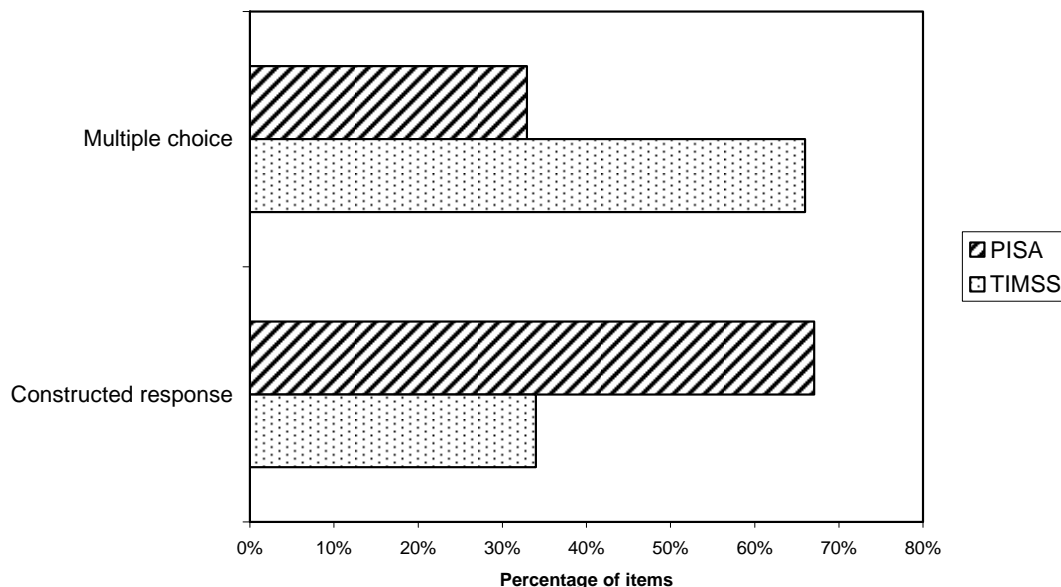
Figure 4.2 focuses on the content domains. Using a classification of PISA items derived by Neidorf et. al (2006, 78) we have been able to classify the PISA items into TIMSS categories. TIMSS has a larger proportion of items focusing on Algebra and on Number, and a smaller proportion focusing on Data and Uncertainty.

**Figure 4.2: PISA and TIMSS 2003 Mathematics Content Domains**



PISA and TIMSS items types are compared in Figure 4.3, with just two categories.


**Figure 4.3: PISA and TIMSS 2003 Mathematics Item Types**



The contrast is quite marked here, with PISA favouring constructed response items by two to one, and TIMSS having the same ratio in favour of multiple choice items.

This rather simplistic set of comparisons has led us to some interesting conclusions:

- TIMSS emphasises items which require the reproduction of facts or standard algorithms, while PISA majors on items which demand connections between existing knowledge.
- TIMSS has a larger number of items focusing on number and measurement, while PISA items are more evenly spread across their content domains.
- A majority of TIMSS items are multiple choice, while a majority of PISA items are constructed response.

Taken together, these observations seem to imply that PISA items are testing deeper mathematical skills, with a requirement for more joined-up thinking and the construction of extended responses. This may be related to PISA's focus on skills for future life rather than grasp of the school curriculum. 

A more in-depth investigation of these issues was carried out by Ruddock *et al.* (2006), who carried out detailed comparison of mathematics items from PISA, TIMSS, the English national curriculum tests at age 14, and the General Certificate of Secondary Education (GCSE) examination taken at age 16. With respect to PISA items, they conclude:

*It is the quantity of reading that marks PISA out, not the complexity of the language, which is similarly unfamiliar in both the international studies. The high reading demand of questions in PISA is often accompanied by a relatively lower demand in the mathematics or science required. This reflects the lower level of mathematics or science that students can apply in new contexts as opposed to very familiar ones. (Ruddock et al., 2006, p.123).*

## **Section 5. Use of IRT**

As is generally the case for international studies, pupils' test results are analysed using Item Response Theory (IRT) - see, for example, Thissen & Wainer (2001) or Hambleton (1991) for a description of such models. However, within this general paradigm there is a certain degree of difference between the two surveys in terms of the type of IRT model used, the number of different scales derived, and how each scale was linked back to the results of previous surveys.

In TIMSS grade 8 2003 (see Gonzalez, Galia & Li, 2004):

- A 2/3-parameter IRT model was used, incorporating a slope parameter as well as item difficulty, and with a 'guessing' parameter included in the case of multiple choice questions. The program PARSCALE was used to fit the data.
- Separate calibrations were carried out, by grouping items appropriately, to generate five mathematics and five science scales. The mathematics scales were: number, algebra, measurement, geometry and data.
- The results were placed on a metric defined by setting the grade eight results for the countries participating in the 1995 survey to a mean of 500, standard deviation 100.

In PISA 2003 (see OECD, 2004):

- A 1-parameter generalised 1-parameter logistic model, referred to in the report as a Rasch model, was used, fitted using ConQuest software.
- A total of seven scales were created: one each of reading, science and problem solving, and four for mathematics (quantity, space and shape, change and relationships, and uncertainty).
- The results in reading and science were scaled to the metric defined in 2000, but for problem solving and mathematics new scales with mean 500 and standard deviation 100 were defined.

We have highlighted the technical differences between the surveys above; however, there is also a substantial amount of agreement in the techniques used. For example, both surveys generate 'plausible values' for each student on each scale, to represent the underlying uncertainty or measurement error involved. Typically five randomly-generated values are produced for each, based not only on the IRT model fitted but also conditioning on a range of background characteristics. In principle, secondary analysis of pupil results from these surveys should be based on such plausible values rather than assuming a single outcome for each individual.

In this area we need to address two questions:

- To what extent do the different modelling assumptions in PISA and TIMSS affect the comparability of results from the two studies?
- How sensitive are these study results in general to modelling assumptions?

Brown, Micklewright, Schnepf and Waldmann (2005) have attempted to address the first issue, comparing results for four surveys including the 2003 PISA and TIMSS studies. The authors were able to compare results for the 1995 TIMSS study using

both a 1-parameter and 3-parameter model, and concluded from this that cross-country patterns of central tendency are robust to choice of model. However, they concluded that this was not true for measures of dispersion, particularly for less developed countries. This does not directly answer the first question above, but it gives us a reason to suspect that some of the different results between surveys may be conditioned by different modelling assumptions.

To answer the second question above we would need to do something rather more radical than has so far been attempted. We have highlighted some of the differences between PISA and TIMSS in terms of modelling outcomes, but in fact the similarities between them are even more striking. Both are deeply committed to IRT analysis, and there are strong similarities in the derivation of pupil scales and the use of plausible values. There are those who are sceptical of the use of IRT – how would we convince them that the substantive findings of these studies are robust and not artefacts of this particular modelling paradigm?

Some sensitivity analysis has been carried out on the impact of changing the details of the modelling on the reported outcomes. The first round of TIMSS was analysed using the 1-PL model, while 1999 and subsequent sweeps were analysed using 2- and 3-parameter models. The results of the 1995 survey were also later re-analysed using 2- and 3-parameter models. The mean scores on a country basis were found to be essentially the same with the exception of one country. However the dispersion of pupil results does appear to be sensitive to the type of IRT model used.

More generally, the 2- and 3-parameter models are dependent on rather strong assumptions, and the more restrictive 1-parameter *a fortiori*. Three important assumptions are unidimensionality, local independence, and universality of application. The results, at least theoretically, are liable to be unreliable if the assumptions do not hold. The analysis teams counter such doubts by extensive testing of the items, but there are still areas of uncertainty.

The first concerns the specifics of testing items. Item fit is tested at the standard .05 significance level, which is not really appropriate for this type of study. The normal use of the .05 level is in testing findings: when you spot an apparent difference in your results, you want to be absolutely sure that it is not due to some kind of chance, and you therefore set a criterion that means that it would have been very difficult for the observed results to have arisen did the difference not exist. Here, it is likely that differences do exist, and we want to be quite confident that they do not. This means in effect that we should set a higher P-value as an item rejection criterion, based on prior calculations of an acceptable level of confidence that an item is fitted by the model.

The second major question is over the question of dimensionality. The analysis of both surveys assumes that the items fall on a 1-dimensional continuum. From a theoretical curriculum-oriented view, this is obviously not the case. Algebra is not Geometry, Biology is not Physics, and so on. See also Schmidt, Houang and McKnight (2005) for a detailed discussion of this. It may be acceptable in mathematics, where the subject is generally compulsory, and ‘marching in step’ within classes may be sufficient to ensure that everyone learns much the same thing at the same rate. (Even within mathematics, there was some evidence in the APU for

differential rates of change between subjects over time.) It is obviously going to be less justifiable within science which consists of a range of much more obviously different activities, which are not unidimensional theoretically, and which may even in some situations be alternatives. Some kind of a theoretical justification for cramming items into a single dimension is required here, especially since the studies then produce subscales.

The number of items used for an assessment is something that requires justification. TIMSS mathematics contained 194 items, 58 for Number, 49 for Algebra, and rather fewer, 29 each for Measurement, Geometry and Data. PISA on the other hand contained between 20 and 23 in the four main categories for its assessment of mathematics as main subject in 2003. For Science for 2003 there were 35 items in total, and the amount of independent information obtained was actually likely to be substantially smaller than this since there were only 13 distinct question stems. (OECD 2004, 419)

A similar question arises with regard to local independence, especially since as noted, PISA's concentration on 'realistic' questions means that a relatively small number of topics can be assessed.

## **Section 6: Collection of Background Information**

Both TIMSS and PISA collect information on the pupils' background as well as school data, while TIMSS also collects information on teachers. In addition, TIMSS views curriculum aspects as one of the main planks of the study, and collects data on this, though this aspect is not discussed here to any great extent.

The rationale given for collecting background information is rather similar in each case.

*'For a fuller appreciation of what the TIMSS achievement results mean and how they may be used to improve students learning (..) it is important to understand the contexts in which students learn'. (Mullis et. al, 2007 P81).*

*'Similarly, it was important to ensure that the framework for contextual factors addressed policy relevant areas and provided the basis for producing internationally significant research' (TIMSS, 2007).*

*'OECD/PISA (...) provides insights into the factors that influence the development of the skills at home and at school, and examines how these factors interact and what the implications are for policy development.'* (OECD, 2003 P10).

Neither actually uses the word 'cause' or 'causation' in their justification, though PISA perhaps comes rather closer than TIMSS. However there is a clear implication that:

- there are school and pupil background factors which affect learning and
- information is being collected in order to be able to assess the effect of these, and how changing them could affect attainment.

*Comparisons between PISA and TIMSS*

A high-level overview of the topics covered in the questionnaires is provided in Tables 6.1 and 6.2.

**Table 6.1 Student and Teacher Questionnaires**

	<b>TIMSS</b>	<b>PISA</b>
<b>Pupil</b>	Pupil and home background	Pupil and home background
	Attitudes to school and school climate	Attitudes to school and school climate
	Attitudes to mathematics	Attitudes to mathematics
	Mathematics activities in school	Mathematics activities in school
	Activities outside school	Learning activities outside school
	Self-rated cognition	Self-rated cognition
		Learning strategies and preferences
		Classroom climate
	Computer use, etc	Computer use, etc
<b>Teacher</b>	Academic preparation and certification	
	Teacher assignment	
	Teacher induction	
	Professional development	
	Teacher characteristics	
	Curriculum topics taught	
	Class size	
	Instructional time	
	Instructional activities	
	Assessment and homework	
	Computers and internet use	
	Calculator use	
	Emphasis on investigation	

**Table 6.2 School Questionnaires**

<b>TIMSS</b>	<b>PISA</b>
School characteristics (location, enrolment, economic background, general atmosphere)	School characteristics (location, enrolment, grade repetition, economic background, instructional hours, limitations to capacity to provide instruction)
Role as principal	Responsibilities in the school
Parental involvement	
Eighth grade instruction in mathematics and science	Instructional organisation for 15-year-olds in mathematics
Eighth grade teachers in your school	Information on teaching staff
Student behaviour	Student attitudes
Resources and technology	Funding, computers
	Admissions criteria
	Assessment
	Options for students with other languages

## Comparisons between PISA and TIMSS

Terminology is different, as always, between the two agencies. However, beneath this we find that there is a fair degree of similarity. Thus TIMSS Home Background corresponds with PISA Student Background, and both seek information on such topics as gender, parental education and employment. The obvious difference between the two is obviously the Teacher Questionnaire, administered by TIMSS but not by PISA. TIMSS questionnaires cover both mathematics and science, while PISA deals with only the 'lead' topic in each sweep. PISA instruments are also somewhat longer, and correspondingly typically more detailed. Perhaps the main difference however lies in the approach to factors hypothesised to influence learning.

These differences correspond to the frameworks of the two studies. Thus TIMSS concentrates to a large extent on the activities of the schools and teachers in classrooms, and on curriculum. PISA's focus is on individuated characteristics of learners - attitudes to the subject, and learning strategies and motivation. Both surveys include a section on problems which make teaching and learning difficult - poor ethos and availability of resources. Interestingly perhaps none of the questions deal with pluses. "My task is made easier by helpful and knowledgeable colleagues and enthusiastic pupils" is not a statement that appears. It might be interesting to consider what this says about the attitudes and assumptions of education professionals worldwide.

The problem about this kind of obsessional data collecting is illustrated by an example from TIMSS. We chose TIMSS results as we are more familiar with them but we are sure that this could equally well be shown in PISA results.

*One of the major ways that students can consolidate and extend classroom learning is to spend time out of school studying or doing homework in school subject. To summarize the amount of time typically devoted to homework in each country, TIMSS constructed an index of out-of-school study time (OST) that assigns students to a high, medium, or low level on the basis of the amount of time they reported studying mathematics, science, and other subjects. On average internationally, and in most of the countries, students at the low level of the index also had lower mathematics achievement, on average, than their classmates who reported more out-of-school study time. However, spending a lot of time studying was not usually associated with higher achievement. On average internationally and in many countries, students at the medium level of the study index had average achievement that was as high as or higher than that of students at the high level. This pattern suggests that, compared with their higher-achieving counterparts, the lower-performing students may do less homework, either because they simply do not do it or because their teachers do not assign it, or more homework, perhaps in an effort to keep up academically.*

(TIMSS 1999 Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade, P125)

To deconstruct this discussion:

1. It is well known that doing homework improves performance
2. The more homework the pupil does, within reason, the better (s)he performs. This is not stated directly but is implied by the need to explain away the lack of a regular trend in the results.
3. We do not find this, at an aggregated level anyway.

4. We shall over-rule this apparent finding by arguing the existence of more than one group of homework takers, one who perform well because they do homework, and one who do homework because they don't perform well.
5. There is no evidence in our findings for this. It just comes from our own experience.

While this is a perfectly reasonable interpretation, and indeed one we would probably have made ourselves had we been forced to express an opinion of such results, it does very strongly question the relevance of this information. How much has been learned here? In this case the writers have sensibly refrained from drawing strong conclusions from the results of their analysis. However, there are many more analyses, especially at national level, where quite unjustified inferences are made from the kind of cross-sectional data available from TIMSS/PISA, though we are not going to embarrass them by referring to them directly.

The point being made here is that there are many examples where an apparent cross-sectional relationship either ignores a common cause, or forgets that causation can be two way. Thus, in the UK pupils in grammar schools perform better on average than those in other types of schools, but this is largely because they are selected to be high performers. Pupils who like mathematics and use certain learning strategies are likely to do well in mathematics but it is also the case that pupils who do well in mathematics are liable to like mathematics more, and to use different learning strategies. Schools that spend more time on task are likely to have higher scores, but conversely teachers in a class with well behaved and high performing pupils are likely to find it easier to spend the time on task, rather than on concentrating on low grade control activities. It is always very difficult, and often impossible, to sort out such tangles of influence.

If we are aware of 'third factors' which we believe influence performance, then a longitudinal study, collecting information on possible third factors would enable far better answers to research questions. Another point is that the major influence on pupils' performance in these surveys will be prior attainment at an earlier point. The apparent influence of school, class and pupil background factors can change quite significantly when allowance is made for prior attainment, in what is sometimes known as a 'value-added' analysis. While the logistic difficulties of longitudinal data collection should not be ignored, the gain in terms of worthwhile data would be considerable.

The problems involved in longitudinal data collection in international studies are exacerbated by the need to provide a system which caters for different educational and data collection structures within individual countries. One possibility would be to try to capture (a subset of) the grade 4 pupils in TIMSS and match those results to their performance at grade 8. One problem is that the time interval between surveys is currently three years rather than four; another is that the grade 8 sampling would need to be modified to ensure that the schools to which these grade 4 pupils transferred were included. However, the major difficulty is probably that of actually tracking pupils through different systems and getting a reasonable number for whom data is available at both time points.

In some educational systems (e.g. England) detailed pupil-level information is available nationally and this tracking process could be carried out centrally and with (in principle) little effort. In other systems, this is not the case and more detailed work would need to be done at a local level. For example, for each grade 4 school it should be possible to determine which were the main schools to which pupils transferred by the grade 8 stage. Lists of names could then be supplied to those schools, who could be asked to identify those who are currently in their grade 8. In this way, with a non-trivial amount of fieldwork, a good proportion of the grade 4 pupils could be tracked longitudinally.

An alternative scheme would involve identifying the main grade for the start of secondary schooling for the majority of interested countries and running a special 'baseline' assessment for pupils in that grade, to be followed up a year or two later with the main study and linked assessments. However it was done, the possibilities for more informative analysis on a significant number of pupils for whom progress over time can be linked to background factors are extremely attractive, and the authors feel strongly that these possibilities should be pursued vigorously by the international agencies and governments in order to add value to the work being done already.

## **Section 7. Survey of Countries Involved**

Many countries had highly similar rankings in TIMSS and PISA: Japan, Korea and Italy, for example. Others had distinctly contrasting ratings, for example New Zealand, Hungary and Russia comparing TIMSS 1999 with PISA 2000, though differences appear to have been less noticeable in 2003. We decided to follow up why some countries had substantially different rankings in the two studies.

We were interested to try to find out whether there was any obvious reason for this, and in particular to see whether there was any **intrinsic** reason for the differences i.e. one relating to the content of the testing, or whether it was due to other factors, such as the definition of the population or details of sampling or administration. One obvious such possible difference lies in the definition of the population, with TIMSS sampling by grade, and PISA by age, but there are many other such possibilities.

For this reason, it was decided to contact coordinators in countries which had participated in TIMSS 1999 and PISA 2000, or in TIMSS and PISA 2003. Those who are familiar with international studies may not be surprised that there is some doubt about what constitutes a country, or indeed 'taking part'. Thus for example, Canada took part in PISA, but only some provinces took part in TIMSS. England took part in TIMSS 1999, the United Kingdom in PISA 2000, England and Scotland in TIMSS 2003 and the United Kingdom in PISA 2003, though the sample in England was judged not to be of sufficient quality to be reported, while the other three constituent countries, including Scotland, were reported individually<sup>5</sup>.

From the websites of TIMSS and PISA, we were able to identify the countries participating, and then the overlap. For convenience, we refer to the 1999 or 2000 surveys as Sweep 1, and the 2003 surveys as Sweep 2, though of course there was a

---

<sup>5</sup> Scotland and Northern Ireland had national reports. Wales did not have a separate sample, so there were only a very small number of schools. In the international report, the overall results for Scotland, Northern Ireland and (erroneously) Wales were reported in one of the annexes.

TIMSS survey in 1995. We e-mailed coordinators from twenty-three countries and received replies of varying degrees of extensiveness from twelve of these, so any statistical description of replies should be taken as indicative rather than representative. There were two main strands to our enquiry. First, we sought any publications comparing the two programmes. Second, we asked a number of questions, mainly about the administration and design of the two studies.

While individual respondents gave careful and extensive answers to the questions, and many countries produced descriptions of the import of TIMSS or PISA results to their schools, there was a surprising lack of officially commissioned comparison of the results of the two studies. England (Ruddock et. al, 2006) produced a comparison between TIMSS and PISA items and those in the public examination and assessment system, and the US produced documents in which the TIMSS items and the PISA items were separately compared to the NAEP (National Assessment of Educational Progress) ones (Neidorf, Binkley, Gattis and Nohara, 2006; Scott, 2006). Ireland also produced a comparison report (Shiel, Cosgrave, Sofroniou & Kelly, 2001), mainly looking at PISA 2000 with references to TIMSS 1995, but we have found little else published in this area.

We now go on to summarise the main replies to our survey, under the headings related to the different questions asked.

***Were there any differences in your country between the results of TIMSS and PISA?***

Of the 13 respondents, three considered that their PISA results were better, one that TIMSS results were better, four reckoned that there were no or slight differences, one stated that there were differences without specifying what they were, and one was unable to answer the question as TIMSS did not survey the entire country. One country reckoned that TIMSS had shown a decline in performance and PISA had not, and one offered the comment that while there was no overall difference, the boy/girl differences were larger in TIMSS.

***If there were any such differences, were you surprised by them?***

Only one country stated that they were surprised by the differences.

***In your opinion, were differences in the samples likely to give rise to differences in the results?***

None of the respondents expressed the opinion that differences in the samples were likely to give rise to differences in the results - except that they obviously related to different age groups.

***In your opinion, were differences in the instructions to schools likely to give rise to differences in the results?***

Of the ten countries that replied, none considered that differences in the instructions to schools were likely to give rise to differences in the results. Not all coordinators were familiar with the details of instructions for both studies.

***In your opinion, were differences in the response rates likely to give rise to differences in the results?***

None of the respondents considered that differences in the response rates were likely to give rise to differences in the results.

To summarise, the main finding from this brief survey is that most respondents considered it unlikely that crude differences in the administration details gave rise to any observed discrepancies between TIMSS and PISA. Many of the differences appeared to arise as a result of the age definition of the population, either directly, or because of curriculum-age interactions, or because of differences in repeating grades. In some countries it was suggested that performance comparisons were in some way biased because operationalisation of the grade definition led to differences in the age of the tested population. Thus, for example, as a result of the grade definition details, the average age of pupils tested for TIMSS 1995 in Scotland was lower, and in some cases up to a year lower, than that of pupils in other countries, for example Germany. This was not the case in PISA, since it had an age-based sample, though conversely it would mean that some students were likely to have had more time at school.

In some countries it appeared that there was an interaction between age and curriculum effects. Thus, in Belgium (Flanders) it was suggested that at grade 8 (TIMSS) Flemish pupils are not taught chemistry and geology, while by grade 10 (PISA), most have been introduced to chemistry, and the topics tested in geology.

The effects of specific examination or testing programs at various ages were also put forward by some of the coordinators as a reason for differences in TIMSS/PISA rankings. Thus Hungary has the entrance examination for upper secondary school at the end of grade 8, and it is considered that the exam awareness arising is likely to raise student achievement in TIMSS. Conversely, the absence of exams may also be a factor. According to school coordinators in 2002, TIMSS 2002/03 was the first formal assessment of a long duration that some New Zealand students had seen. Similarly it is considered that Scottish pupils may be less used to formal assessment than students in other countries, since there are no National tests or new National Assessments in science in Scotland.

As intimated above, national policies on repeating years differ between countries. At the one extreme, England's pupils mainly move up through the schools by age, to such an extent in fact that there are not any statistics about this. In other systems, there is a sizeable amount of grade repetition, so that in Flemish-speaking Belgium, for example, 72.2 per cent of those tested were in grade 10, 22.8 in grade 9, and 2.5 per cent in grade 8.

This discussion has looked at obvious structural factors which could confuse inferences based on the overt performance data on TIMSS and PISA. We now turn to a direct comparison between those results. It is important to bear in mind that differences in scores between TIMSS and PISA do not in themselves represent differences in absolute performance. Even were the two tests comparable, the PISA sample is substantially older than the TIMSS sample, although each test is standardised to an international mean of 500. The plots should give an impression of

*Comparisons between PISA and TIMSS*

the relative ordering of countries. Results have been reported in original country scores, rather than rankings, to establish the relative sizes of differences.

Figures 7.1 and 7.2 show the mean PISA scores graphed against the TIMSS scores for the countries doing both in 2003. Parts of this are not particularly easy to see, so Figures 7.3 and 7.4 show the standardised residuals when predicting PISA scores from TIMSS.

**Figure 7.1: TIMSS vs PISA 2003 Mathematics**

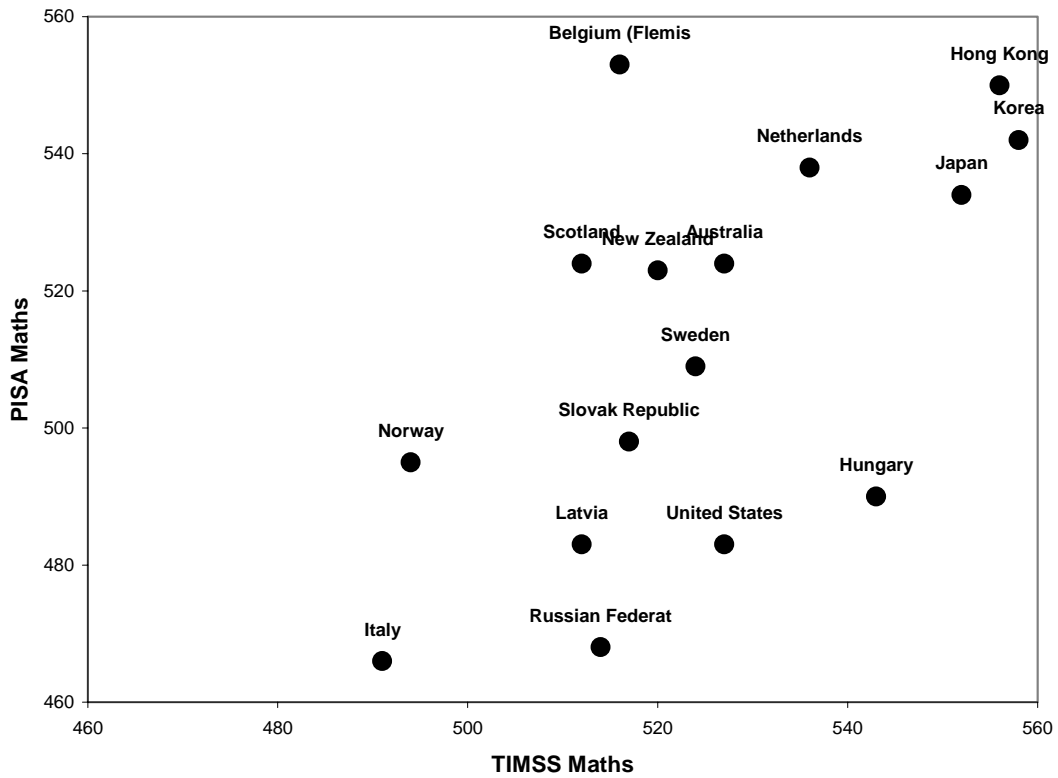


Figure 7.2: TIMSS vs PISA 2003 Science

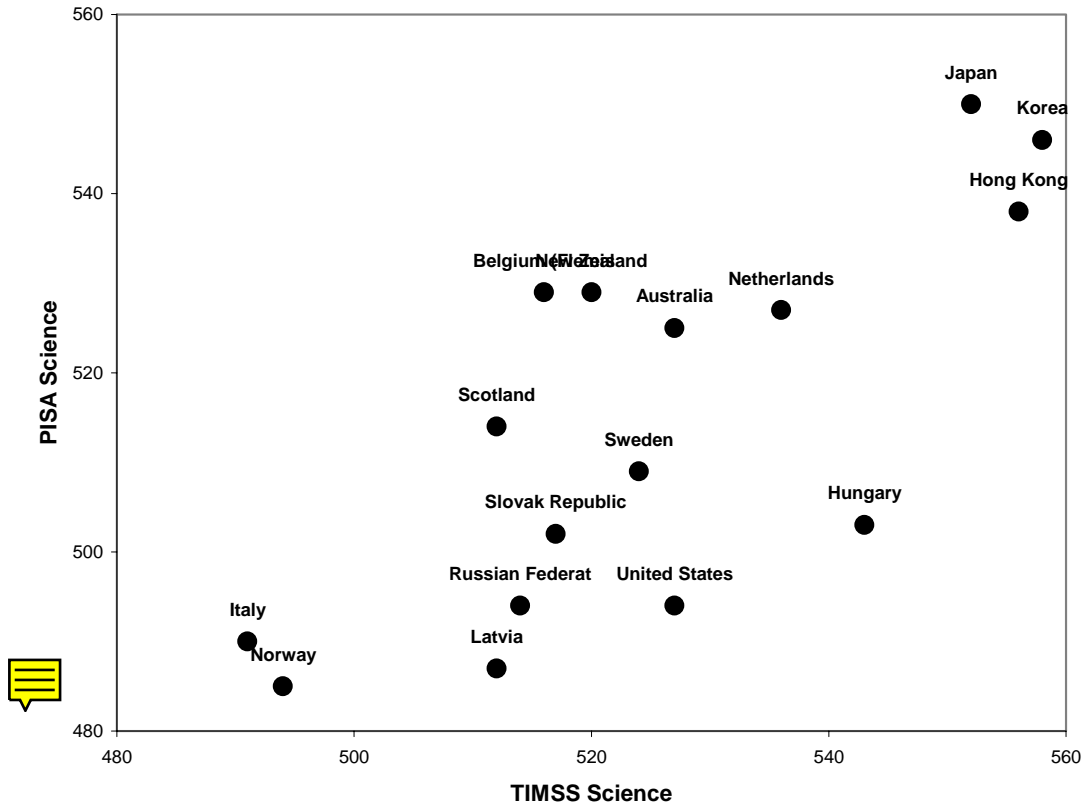


Figure 7.3: PISA Maths 2003 Residuals (from TIMSS Maths 2003) vs TIMSS

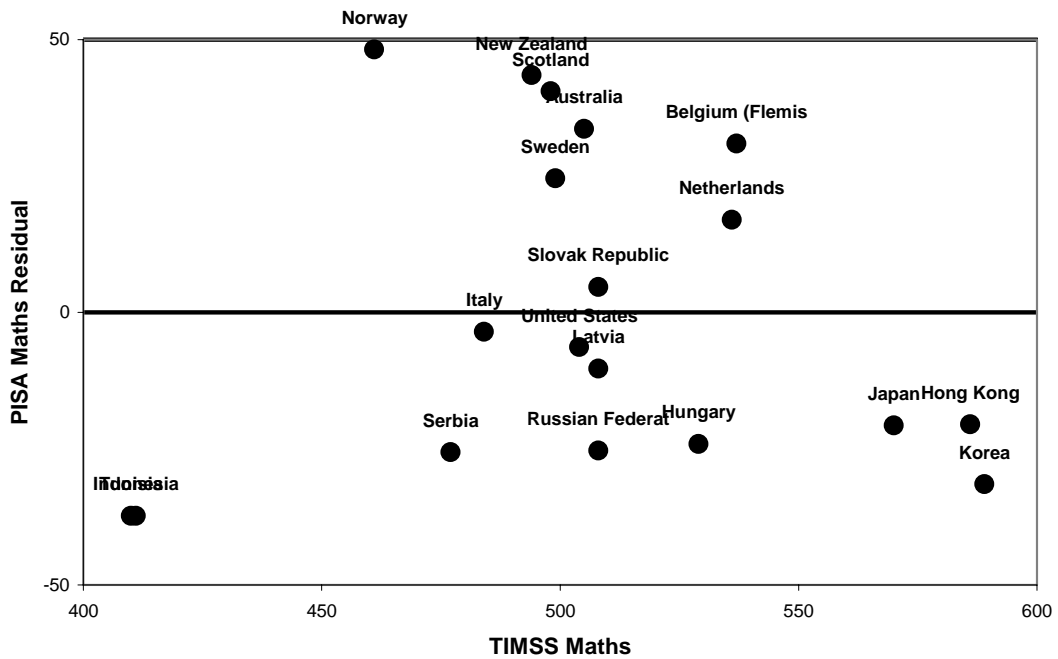
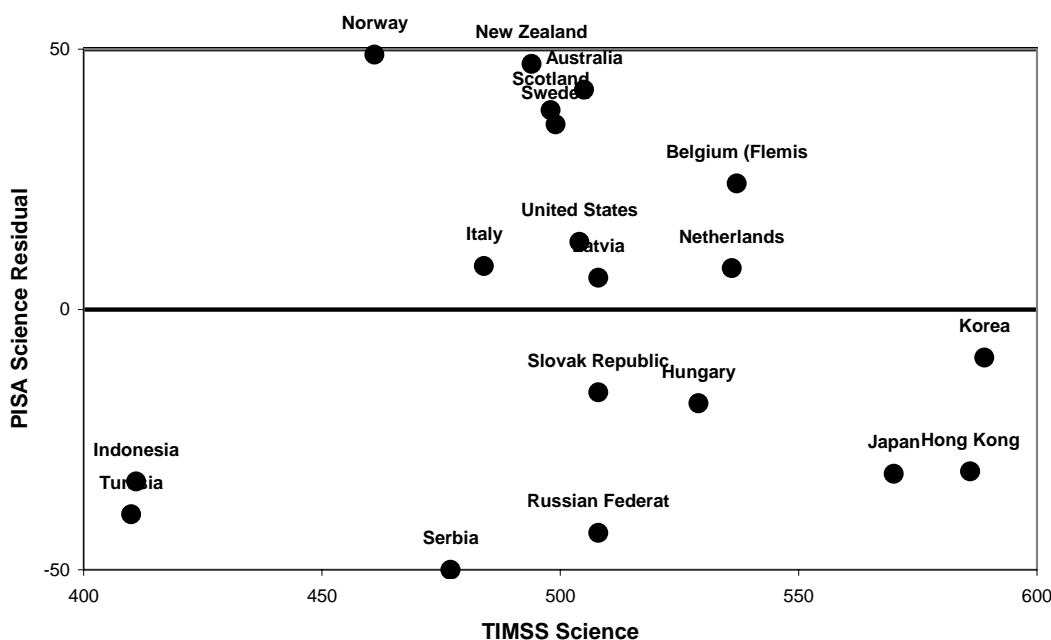


Figure 7.4: PISA Science 2003 Residuals (from TIMSS Science 2003) vs TIMSS



Bearing in mind all the caveats and possible alternative explanations just described, there is a clear suggestion of a pattern, with the ‘first world’ countries doing better, and former ‘Eastern bloc’ countries worse, on PISA than would be predicted on the basis of their TIMSS performance. Interestingly, the Hungarian national coordinator had this to say about the Hungarian situation:

*‘TIMSS focuses on the curriculum related tasks, while PISA is literacy-based. Hungarian school system is still highly relies on factual knowledge and traditional teaching strategies, so students are relatively good in tasks which are close to their usual classroom tasks, while they meet relatively few literacy-based tasks and they do not know what to do with these’.*  
( Balazsi, 2006)

Comparable results emerge from the TIMSS 1999- PISA 2000 comparison. It seems that the PISA assessment may be more aligned with the education systems of the most highly developed countries than is TIMSS.

## Section 8. Conclusions

Do we in fact have two distinct exercises? The rhetoric suggests that we do. We have mathematics literacy instead of mathematics, and science literacy instead of science. There is an apparent difference in the philosophy of approach. Barry McGaw (2002a), the Director for Education of the OECD, characterised the difference as TIMSS being interested to discover, “what science have you been taught and how much have you learned?”, while for PISA it was “what can you do with the science you have been taught?”. If one can sum up the differences in the images of the two studies, TIMSS is inside the school wondering what makes it tick, while PISA is outside in the street waiting to see what’s coming out. There appears also to be a more individual- oriented approach to the data: where TIMSS has detailed

investigations of what is taught and how, PISA's angle is to compare the learning styles and motivations of the individual pupils.

So does this conform to an 'on-the-ground' difference between the two studies? If they were very similar, then simply putting different labels on them would not make them different. As has been covered earlier in this paper, there are in fact a very large number of similarities between them. In fact, the PISA sampling and analysis appears to have been largely lifted from the TIMSS model, built up as it has been over years of experience. To take just a few examples at random, both studies took about 4000 pupils, using a pps sample of schools, used a matrix sampling design, and analysed the data using IRT and plausible values. We could go on. To be sure, PISA analysed at a different stage of education, took a year group rather than a grade group, sampled classes rather than individual pupils, and used the 1-parameter model rather than a 2- or 3-parameter model. These are not essential differences, though they may give rise to some alterations in the rank ordering of countries, as discussed earlier.

The main difference lies in the type of item. PISA items are aimed at 'life skills', while TIMSS items are more knowledge-oriented. However, to quote Smithers (2004), "Just because it says it's curriculum-free, doesn't necessarily mean it is". Where TIMSS questions are more direct and abstract, PISA questions are more lengthy and wordy. TIMSS asks 'What do you know?' while PISA asks 'What can you do?'

Some differences in results appear to be based substantially on the differences in the interaction between age and curriculum, with, for example, Flemish Belgium appearing to perform relatively less well at age 13 (TIMSS) because they have not generally been introduced to algebra but catching up by age 15. However we appear to have established by our small survey of coordinators that otherwise sample definition, response rates or other technical questions do not appear to affect the national scores differentially between the two studies. Even after allowing for possible alternative explanations there seems to be a pattern that highly developed 'Western' countries do relatively well on PISA, while the former Eastern Bloc countries do less well. This corresponds too, as far as we can see, with how the countries concerned view their own systems, with the Eastern Bloc countries concentrating on a more formal 'traditional' approach and the 'Western' countries going for interpretation and application.

Thus, there do seem to be some differences, and probably ones that reflect useful aspects. It is useful to know if the education system is 'working'. It is also useful to know what the graduates of the system can actually do. However, if the two studies are supposed to be different, should they not really be a bit more different than they actually are? 'Life skills' are presumably the ability to hold down a job, to engage in basic financial dealings, and to be able to make political judgments based on intelligent and informed assessment of the claims of politicians. If this is what PISA is supposed to be measuring, why is it conducting its assessments in schools, at a stage some way before most pupils leave education, using a paper and pencil test administered to pupils sitting in rows in classrooms? And what evidence is there that these really are life skills?

So, we have two comparative education survey systems – but does the world actually need them both? Perhaps before asking this, the first question to ask is, does the world

actually need one? The first and most noticeable outcome of these studies is the 'league table' of countries, which always nowadays creates a stir, especially if the country concerned has been unexpectedly unsuccessful. In our opinion, and that of experts in the area, too much is made of the league table rankings. To be fair, the studies themselves are aware of this, and appear to concur with our view. PISA, for example gives the ordering of its results in statistically comparable groups, though in our opinion these are based on estimated standard errors which are too small as they take no account of item sampling, a potentially serious problem in PISA with its relatively small number of items. There certainly is a potential problem of over-interpretation, and particularly biased interpretation, of the results. Smithers sums this up in a delightful sentence:

*'The cloud of data generated becomes a canvas on to which the committed can project what they want to see.'* (Smithers, 2004).

In a modern society, where governments feel it necessary to try to 'run' the way society functions, it seems to be useful to make sure that you have the right information, and international studies do offer the possibility of obtaining comparative information that would not be obtainable in any other way. For example, for the UK this is the only internationally comparable source for the output of the education systems within the country<sup>6</sup>. More generally, it gives countries trying to plan their future an indication of strengths and weaknesses in their skills base.

TIMSS and PISA both cost participating countries of the order of \$2m per study. Perhaps more to the point is the level of demand, especially at secondary school level on already overburdened schools. It is a point of interest that the very first piece of education research done by one of the authors was in investigating an accusation that schools were overloaded with requests for statistical information, a complaint sparked off by none other than one of the first IEA studies. *Plus ça change*.

Examination and testing is a social product, and as such evolves with change in society. Up to and immediately post-war, western society was hierarchical, rigid and unchanging with relatively few desirable occupations. There was relatively little change in the information available, and the main aim of testing was to order the candidates on their skills and use this to guide and justify entry into various activities. More recently, with the accelerating pace of development of knowledge, and the widened range of opportunities the focus has shifted to ensuring the maximum effective take up of particular skills. With the coming of globalisation, outsourcing and computerisation, the nature of employment has largely changed, so that in developed countries it is more important to be able to analyse a situation and specify what is to be done, than to be able to actually do it. In parallel there has been an extensive development in the democratic process, or at least the ability to make an input into the running of society, in a very large number of countries over the world. And this in turn points to the importance of being prepared to assess to some extent the validity of the statistical basis of political arguments. This implies that the PISA approach of aiming to measure literacy in the areas concerned is an important one.

---

<sup>6</sup> Or at least it would be if they all took part successfully in the same study at the same time.

Both TIMSS and PISA set themselves up to measure change over time, the former since 1995 and the latter since 2000. Indeed, with TIMSS, the name Trends in Mathematics and Science Survey could be taken as a bit of a clue to this orientation. While this is a strength, it is also a source of vulnerability, and a study can become a prisoner of its past. It is well known that if you want to measure change, then do not change the measure. The measure cannot remain the same because at least some of the items are released after each sweep. The use of the Rasch or IRT models is aimed to deal with this. However, even if it is accepted that this is possible, it can only occur if the tests are measuring ‘the same thing’. Conceptually and in practice the PISA items have a rather different orientation to those in TIMSS. Had TIMSS wanted to move to assessing life skills, it would have been difficult to link this to the trends in measures up to that point. And PISA will undoubtedly face the same challenge in the future when the global agenda moves on. How do such studies cope with maintaining links with what has come before, while at the same time ensuring that test practice and orientation remains as up to date as possible? Do the studies keep adding on new bits, eventually becoming so unwieldy they sink under their own weight? Or do they simply jettison the old parts and lose the contact with the past? In our opinion the need for comparability over long periods of time is less important than one might imagine. At present there are so many changes in curriculum and emphasis that time series much over about ten years are really not very informative.

### ***Questions***

- The two studies differ in the intensity of the coverage and the number of items, with TIMSS having over twice the number of items than in a ‘main’ PISA year, and five times the number in a ‘minor’ PISA year. The effective number of items is actually rather smaller because of the PISA habit of embedding a number of questions within a single question stem. While this improves realism, it also potentially reduces generalisability. Can PISA justify an item sample this small, and are the ‘minor’ PISA years of any value?
- PISA aims to assess 13 different science topic areas. Is it really feasible to produce valid scale scores in all these areas?
- On the other hand, what is the theoretical (as opposed to the practical) justification for using any kind of IRT model, with its unidimensionality assumption, on such a heterogeneous topic area as science?
- How sensitive are the main outcomes of the studies to the different models used for analysis?
- Would it be possible to introduce some kind of computer adaptive testing to cut down the volume of assessment to get a given degree of precision?
- We were able to find little formal research into the reasons for the apparent differences in relative positions between the two studies. Why is such research not a matter of the highest priority for both agencies?

### ***Recommendations***

One aspect of the apparent competition between TIMSS and PISA is that it shows the importance of constant development both of aims and techniques. It is arguable that PISA has stolen a march on TIMSS by the introduction of the ‘life skills’ area. This

emphasises the importance of maintaining a methodological development function in any such studies.

It seems unfortunate that there are two large international studies in the field at the same time doing very similar things. As such they are bound to view themselves as competing. While competition can sharpen up an outfit's act, it can also have unfortunate consequences. PISA was able to hit the ground running to the extent that it has largely because it lifted its methodology from TIMSS. IEA has always been extremely open in the best academic way in allowing inspection of its methodology, but after this experience one would hardly blame them if they were to shield their working under the guise of commercial confidentiality. This would hold up the development of research capacity in many countries, and also make it less likely that methodology would be tested and proved in the normal academic procedure.

- We recommend that both IEA and OECD maintain open access to their methodologies, and encourage criticism and debate from the wider academic community. This would go some way to dispelling the impression that is sometimes given of a 'closed shop' in this area, and should be done in a spirit of openness and willingness to learn and improve, recognising there is not necessarily a 'correct' answer to each technical problem.
- We further recommend that more analysis be carried out on the sensitivity of the main survey results to the exact details of the models used (e.g. 1-parameter versus 2-parameter IRT). Such sensitivity analyses could be carried out by third parties using the raw data.
- We suggest that the importance of comparing attainment over long time periods is downgraded, to a maximum of 10 or possibly 12 years, since education curricula and aims changes so rapidly that old results are no longer relevant. This may impact on the frequency of studies, and also the extent to which common items are kept in for several surveys.
- It is not possible generally to form a valid impression of the effect of any aspect of school structure or practice, or pupil attitudes or learning styles from a cross-sectional study. We recommend that the feasibility of introducing a longitudinal study, following the same pupils over a period of years, should be actively investigated. While recognising the logistical challenges, we believe the potential benefits, in terms of increased understanding of educational systems, are significant.
- We would recommend that there should be careful in-depth investigations of the apparent discrepancies between TIMSS and PISA results at the country level.

Finally, IEA has formed an enviable reputation for scholarship, research and integrity over its long and distinguished career. It is important for the future of international that this should continue. We believe that this will be best continued by maintaining its independent non-governmental status.

## **Acknowledgements**

We are very grateful to the following for encouragement and helpful comments: Lorna Bertrand, Tom Loveless, Scott Murray, Graham Ruddock, Liz Twist, Chris Whetton.

We should like to thank very warmly all those who responded to our short survey. Not all these wished to be named, but the respondents included Dianne Pennock, Dominique Lafontaine, Inge De Meyer, Ildiko Balazsi, Anna Maria Caputo, Mee-Kyeong Lee, Erna Gille, Marit Kjaernsli, Karl-Goran Karlsson and Megan Chamberlain.

## **References**

- Balazsi, I (2006) Personal communication.
- Beaton, A (ed) (1988) *Expanding the new design: the NAEP technical report*. Princeton, NJ: Educational Testing Services.
- Bertrand, L. (2006) Personal Communication
- Brown, M. (2000) *Problems in Interpreting International Comparisons*
- Brown, G., Micklewright, J. Schnepf, S. and Waldmann, R. (2005). *Cross-national surveys of learning achievement- how robust are the findings?* Southampton statistical Services Research Institute Applications and Policy working Paper A05/05.
- Clarke, C. (2003) Speech to NASUWT Conference, Bournemouth, 24 April 2003.
- Council of Ministers of Education, Canada (CMEC) (2000)  
Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA), and Youth in Transition Survey (YITS)  
FACT SHEET GENERAL FACTS ABOUT PISA).  
<http://www.cmec.ca/pisa/2000/factsheet.en.pdf> 21 September 2006
- Foxman, D., Hutchison, D. and Bloomfield. B. (1991) *The APU Experience*. HMSO.
- Gibb, N.(2002) House of Commons Hansard Debates for 15 Nov 2002.
- Hambleton, R. (1991) *Fundamentals of Item Response Theory* . : Sage.
- IEA (2006) International Association for the  
Evaluation of Educational Achievement Site Info Mission Statement  
[http://www.iea.nl/mission\\_statement.html](http://www.iea.nl/mission_statement.html) 21 September 2006
- McGaw, B (2002a). Paragraph 1, Examination of Witnesses, Select Committee on Education and Skills, 20 March 2002,  
<http://www.publications.parliament.uk/pa/cm200102/cmselect>

Martin, M. O., Mullis, I. V. F., Chrostowski, S. J., TIMSS (2003) Technical Report. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. ISBN: 1-889938-35-1

Miliband, D. (2003a). Speech to ATL Conference, Blackpool, 15 April 2003.

Mullis, I. V. S., Martin, M. O., Graham J. Ruddock, Christine Y. O'Sullivan, Alka Arora, Ebru Erberber (2007) TIMSS 2007 Assessment Frameworks. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College ISBN: 1-889938-39-4

Neidorf, T.S., and Garden, R. (2004) Developing the TIMSS 2003 Mathematics and Science Assessment and Scoring Guides, Chapter 2 in Martin, M. O., et. al (2003) Technical Report. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. ISBN: 1-889938-35-1

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved 22 September 2006 from <http://nces.ed.gov/pubsearch>.

Normington, D. (2002a). 'Transforming secondary education', *Education Journal*, issue 62, June 2002.

OECD (1999) Measuring Student Knowledge and Skills- A New Framework for Assessment.

OECD (2003) The PISA 2003 Assessment Framework

OECD (2004). Learning for Tomorrow's World: First Results from PISA 2003.

Prais, S.J. (2003) 'Cautions on OECD's recent educational survey (PISA)', in *Oxford Review of Education*, Vol. 29, No.2, pp.139-163.

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

Ruddock, G., Clausen-May, T., Purple, C. and Ager, R. (2006) Validation study of the PISA 2000, PISA 2003 and TIMSS 2003 international studies of pupil attainment. Nottingham: DfES.

Schmidt, W. H., Houang, R. T. and McKnight, C. (2005) Value Added Research: right idea but wrong solution? In Lissitz, R. (ed) Value Added Models in Education: theory and Applications. Minnesota: JAM Press.

Scott, E. (2006) *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat*

*Comparisons between PISA and TIMSS*

[http://nces.ed.gov/TIMSS/pdf/naep\\_timss\\_pisa\\_comp.pdf](http://nces.ed.gov/TIMSS/pdf/naep_timss_pisa_comp.pdf) 29 September 2006

Shiel, G., Cosgrave, J., Sofroniou and Kelly, A. (2001) *Ready for Life? The Literacy Achievements of Irish 15-Year Olds With Comparative International Data*, Dublin: Educational Research Centre

Smithers, A. (2004) *England's Education: what can be learned by comparing countries?* University of Liverpool: Centre for Education and Employment Research

Thissen, D. and Wainer, H. (eds) (2001) *Test scoring*. Mahway, New Jersey: Lawrence Erlbaum Associates.

TIMSS (2007) *The TIMSS Assessment Frameworks and Specifications 2003*  
<http://timss.bc.edu/timss2003i/frameworks.html> 14/9/06