Chapter 2-6

A COMPARISON OF MATHEMATICS PERFORMANCE BETWEEN EAST AND WEST: WHAT PISA AND TIMSS CAN TELL US

Margaret WU
University of Melbourne

1. INTRODUCTION

International studies such as the Trends in Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) provide results of relative mathematics achievement outcomes of countries. While the ranking of countries in terms of mathematics performance is interesting for both policy makers and the scientific community, what is more important for the mathematics education community is to identify factors that underlie the differences between countries. More specifically, while we have an interest in overall assessments of how students perform, we are perhaps more interested in understanding the nature of the differences between countries.

In this paper, we report a more focused comparison of Eastern and Western countries in students' performance in the PISA 2000 mathematics assessment. We examine the cognitive demands of each item and identify item characteristics that influence, in different ways, students' performance across the Eastern and Western cultures. In particular, we look for patterns and similarities within a group of Eastern countries and a group of Western countries. However, before delving into the item analysis of the PISA assessment, we first take a look at the differences between TIMSS and PISA results. We observe a pattern of differences that leads us to form a hypothesis about the differential item functioning (DIF) in Eastern and Western countries.

2. BACKGROUND OF PISA AND TIMSS

PISA is an international comparative study conducted by the OECD. The main aim of the project is to assess 15 year-old students' knowledge and skills in a number of subject domains, with an emphasis on these students' preparedness for life (OECD, 1999). PISA is intended to be an on-going study, with data collection conducted every three years. The first cycle of PISA (PISA 2000) spanned four years, from 1998 to 2001, with the main study data collection conducted in the year 2000. Thirty-two countries participated in this survey. PISA 2000 assessed reading, mathematics and science, with reading as the major focus. For mathematics, there were 60 minutes of testing material in the assessment, but only five-ninths of the students were administered mathematics items, and each of these students received 30 minutes of mathematics items in a rotated-forms test design (Adams and Wu, 2002).

TIMSS (Third International Mathematics and Science Study) was an IEA (International Association for the Evaluation of Educational Achievement) study first conducted in 1994-1995, with 41 countries participating at 5 grade levels. TIMSS 1999, also known as TIMSS Repeat or TIMSS-R, is a replication of TIMSS at the lower secondary school level – the eighth grade level in most countries, with an average student age of 14.4 years.

A key difference between PISA and TIMSS is that PISA has a "literacy" based orientation with a goal of "assessing the extent to which young people have acquired the wider knowledge and skills that they will need in adult life" (OECD, 1999). The PISA mathematics framework states that

> The term *literacy* has been chosen to emphasise that mathematical knowledge and skills as defined within the traditional school mathematics curriculum do not constitute the primary focus of OECD/PISA. Instead, the emphasis is on mathematical knowledge put to functional use in a multitude of different contexts and a variety of ways that call for reflection and insight.

TIMSS, on the other hand, starts the development of the assessment framework by surveying the mathematics curricula of all participating countries (Mullis *et al.*, 2001), although the TIMSS assessment framework is not solely based on the overlap of the curricula of participating countries. While the starting points of PISA and TIMSS are different, there is no doubt that the PISA assessment has considerable overlap with TIMSS, as the designers of curricula generally also aim for preparing students for skills needed in their future life. The differences between PISA and TIMSS are mainly in the emphasis of various skills and the manner in which questions are posed, rather than any fundamental differences in mathematics content.

As mathematics was a minor assessment domain in PISA 2000, only two areas of mathematics applications, referred to as *big ideas* in PISA, were chosen for the assessment: *change and growth*, and s*pace and shape*. The PISA mathematics framework gives the following rationale for the selection of these two big ideas:

> First, these two domains cover a wide range of subjects from the content strands. Second, these domains offer an adequate coverage of existing curricula. Quantitative reasoning was omitted from the first survey cycle because of the concern that it would lead to an over-representation of typical number skills.

The fact that PISA avoided the inclusion of purely computational items reflected the general thinking of the expert group that advised on the PISA mathematics assessment. Few PISA items required only the recall of knowledge. Most PISA items focused on "analysing, reasoning and communicating ideas". So the PISA items were largely concerned with the application of mathematical ideas and making sense of mathematics, not only about knowing algorithms or computational procedures. In analysing PISA data and in assessing the differences between the results of PISA and TIMSS, it is important that we bear in mind the differences in the conceptualisation of these two projects.

3. OVERALL MATHEMATICS PERFORMANCE IN PISA

As is generally the case in international studies of mathematics, PISA showed that Asian countries outperformed western countries when mean mathematics proficiencies at the country level are compared. Figure 1 shows the country mean scores and 95% confidence intervals (OECD, 2001). In PISA, only two countries are from East Asia: Japan and Korea. It can be seen from Figure 1 that Japan and Korea outperformed all other countries in mathematics in PISA. In general terms, we summarised the PISA results as follows: Students in Asian countries had the highest average scores, followed by students in English-speaking countries, northern European countries, eastern European countries, southern European countries, and then central and south American countries.

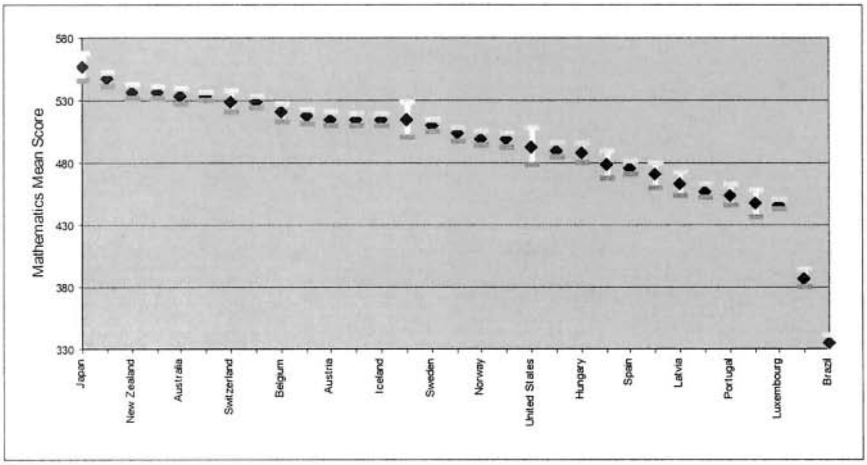


Figure 2-6-1. PISA Mathematics Mean Scores

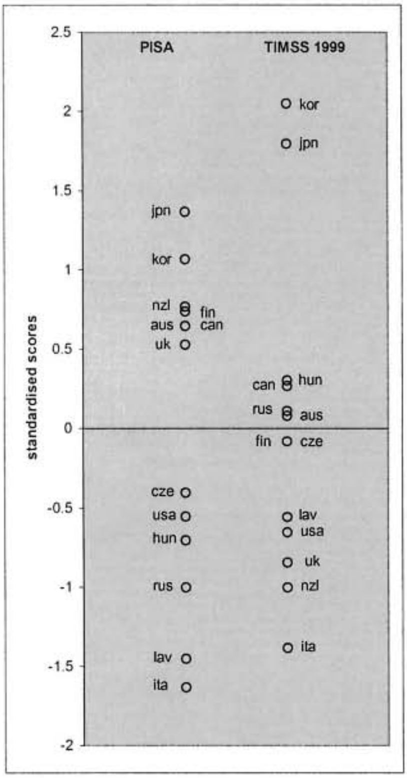


Figure 2-6-2. Standardised PISA and TIMSS 1999 Scores

When PISA 2000 results are compared to TIMSS 1999 (Mullis *et al.*, 2000) results, a number of striking differences can be seen. Figure 2 shows a comparison of PISA 2000 and TIMSS 1999 results. In this figure, we include the 13 countries that took part in both PISA 2000 and TIMSS 1999: Australia, Canada, Czech Republic, United Kingdom, Finland, Hungary, Italy, Japan, Korea, Latvia, New Zealand, Russia and the United States.

The PISA 2000 and TIMSS 1999 results are not scored on the same scale. They need to be standardised to make them comparable. The country mean scores were standardised in the following way. We computed the means and standard deviations of the 13 country means for each study, and computed the standardised score for each country by subtracting the mean and dividing by the standard deviation of the 13 mean scores in each study. For example, a standardised PISA score of 1.37 for Japan means that Japanese PISA mean score is 1.37 standard deviations away from the overall mean of the 13 PISA country mean scores. Similarly, we see that Korea is 2.05 standard deviations away from the mean of the 13 TIMSS 1999 country means. In this way Figure 2 shows the relative standing of each of the 13 countries in PISA 2000 and in TIMSS 1999, in terms of the number of standard deviations from the mean of the group of countries under comparison.

We can make three observations about Figure 2. Firstly, in TIMSS 1999, there is a large gap between the Asian countries (Japan, Korean) and the rest of the countries. In PISA, the gap is narrowed. Secondly, English-speaking countries performed well in PISA 2000 and were not too far behind Japan and Korea. However, in TIMSS 1999, English-speaking countries performed relatively poorly. Thirdly, eastern European countries performed poorly in PISA 2000 as compared to their performance in TIMSS 1999.

Figure 3 displays a scatter plot of standardised PISA 2000 and TIMSS 1999 scores for the 13 countries. This plot shows even more clearly the difference between PISA 2000 and TIMSS 1999 scores between two groups of countries.

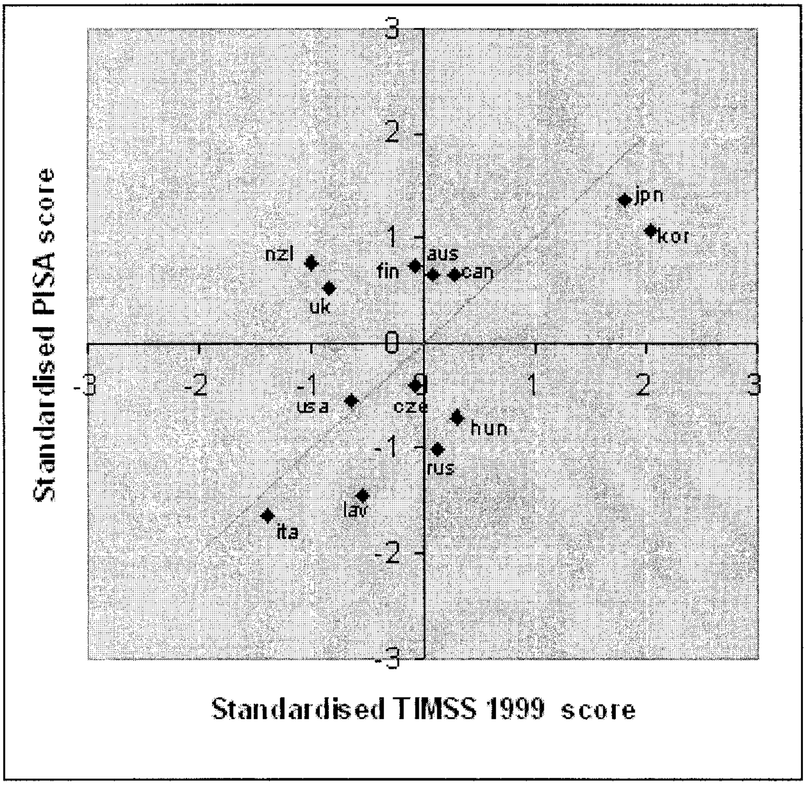


Figure 2-6-3. Standardised PISA 2000 score versus standardised TIMSS 1999 score

A 45° line is drawn to indicate the line of equality between standardised PISA 2000 score and TIMSS 1999 score. If we group Japan and Korea together with eastern European countries and define this as Eastern countries, then it is clear that Western countries are above the line of equality, and Eastern countries are below the line, with the exception of Italy. This means that Western countries performed relatively better in PISA 2000 than they performed in TIMSS 1999, and, generally speaking, Eastern countries performed relatively better in TIMSS 1999 than in PISA 2000. From the point of view of mathematics curriculum design, eastern European countries have more similarities with Japan and Korea than with Western countries, in terms of the content of traditional and formal mathematics taught in schools. Therefore, this finding of the two groups is not surprising.

The rank orders of countries are also quite different between PISA and TIMSS-R. Table 1 shows the rank orders of the 13 countries for PISA 2000 and for TIMSS 1999:

Table 2-6-1. Rank orders of countries in PISA 2000 and TIMSS 1999

	PISA 2000 rank	TIMSS 1999 rank
Japan	1	2
Korea	2	1
New Zealand	3	12
Finland	4	8
Canada	5	4
Australia	6	6
England	7	11
Czech Republic	8	7
United States	9	10
Hungary	10	3
Russia	11	5
Latvia	12	9
Italy	13	13

New Zealand has moved from third place (out of 13 places) in PISA 2000, down to 12th place in TIMSS 1999, while Hungary has moved from 10th place (out of 13 places) in PISA 2000, to third place in TIMSS 1999. Such discrepancies will no doubt raise questions like "which set of results is more *valid*?" and "why are there such differences?" These questions are not easy to answer. We are unlikely to find a simple answer to the question of assessing the validity of the results. But we can uncover some of the reasons for the differences. Three variables might be causes of the differences: population definition, time of the survey, and test content. The target populations in the two surveys are different in that PISA is age-based and TIMSS is grade-based. The average age of the samples of students for PISA is about one year older than the average age of the samples for TIMSS. The PISA survey occurred about one year after the TIMSS 1999 survey. In some sense, PISA 2000 and TIMSS 1999 essentially tested the same cohort of students in the countries, although one may argue that an age-based sample captures a slightly different cohort from a grade-based sample due to factors like retention. It is also possible, but unlikely, that Eastern countries had a program that accelerated students' learning from 14 to 15 years-old. We do not think that the age definition and the time of survey are likely to have caused the differences we observed. In this paper, we will examine the third variable, test content, in more detail. To try and better understand how the test content can affect performance, we need to examine item characteristics of each assessment. Our first hypothesis is based on the conceptual difference between PISA and TIMSS, as described earlier. The main difference is that PISA is not curriculum-based. The mathematics curricula in the 13 countries differ in varying degrees to the PISA mathematics framework. PISA's approach of assessing applications of mathematics may present more challenges to students who are used to learning mathematics in a more formal way. In the following sections, we examine students' performance on different types of items. In doing so, we hope to identify performance patterns that can be related to item characteristics.

4. PERFORMANCE AT THE ITEM LEVEL

To keep the interpretation of results manageable, we start with the analysis of four countries only. As this paper is primarily concerned with a comparison of the East and the West, where East is defined as East Asian countries, we include Japan, Korea, Australia and the U.S.A. in our first analysis. We use PISA 2000 assessment data to carry out this analysis, as we are familiar with the items, having been involved in the test development process. Owing to the embargo on a number of items that are used for linking purposes for future PISA cycles, we are not able to describe in detail all the items used in this assessment. We will illustrate our findings using some released items, which are included in the Appendix.

There were 31 mathematics items in the PISA 2000 database. One item was deleted in Japan owing to translation errors, so we will include only 30 items that were common to all countries in the following analyses. Table 2 and Figure 4 show the estimates of item facilities for these 30 items by country.

Of the four countries, Japan scored the highest country mean score, Korea is the next highest, followed by Australia, then the United States. If all items behave in a similar way in the four countries, we expect to see the percentages correct for each item also following the same order: Japan, Korea, Australia, the United States. For example, item 2 (M034Q01T) shows the four facilities (56.5, 51.3, 43.9, 29) in the order we expect, according to the order of country mean scores. Similarly, items 3, 4, 5, 7, all show the same ordering in terms of percentages correct. The same pattern can be seen in Figure 4, where the percentages correct are displayed visually. The fact that Figure 4 shows percentages correct moving up and down across items reflects the item difficulties of the items. In general, when an item is relatively difficult, the percentages correct are low for all four countries. Similarly, when an item is easy for one country, it is usually easy for all other countries. We see the four percentages for each item moving in relative unison across most, but not all, items.

For some items, we observe that the ordering of percentages correct is not quite as expected. For example, item 27 (M179Q01T) shows that Australia and the United States performed better than Japan and Korea. Figure 4 also shows that for some items, the four percentages correct are far apart from each other, and for other items, the percentages correct are close together. When items exhibit varying relative difficulties across countries, we say that there is Differential Item Functioning (DIF) on these items.

Table 2-6-2. Percentages correct of PISA 2000 mathematics items by country

	M033Q01	M034Q01T	M037Q01T	M037Q02T	M124Q01
Item No.	**1**	**2**	**3**	**4**	**5**
jpn	81.5	56.5	81.6	85.6	46.1
kor	74.0	51.3	70.9	80.0	41.2
aus	77.4	43.9	66.8	63.4	30.8
usa	72.5	29.0	46.4	59.8	25.5

	M124Q03T	M136Q01T	M136Q02T	M136Q03T	M144Q01T
Item No.	**6**	**7**	**8**	**9**	**10**
jpn	37.2	81.5	50.8	21.1	84.9
kor	11.7	73.4	60.6	30.4	78.6
aus	19.9	61.7	25.4	19.3	72.9
usa	17.6	53.4	23.9	14.8	52.7

	M144Q02T	M144Q03	M144Q04T	M145Q01T	M148Q02T
Item No.	**11**	**12**	**13**	**14**	**15**
jpn	41.7	85.8	49.9	72.6	23.3
kor	35.5	78.8	49.8	63.7	15.2
aus	29.6	86.1	43.0	64.6	26.7
usa	12.4	74.3	34.8	52.4	21.1

	M150Q01	M150Q02T	M150Q03T	M155Q02T	M155Q03T
Item No.	**16**	**17**	**18**	**19**	**20**
jpn	76.6	77.5	45	63.5	22.5
kor	77.5	86.7	48.3	68.0	22.9
aus	64.3	73.1	63.7	73.5	19.2
usa	50.8	61.2	57.2	64.2	18.1

	M155Q04T	M159Q01	M159Q02	M159Q03	M159Q05
Item No.	**21**	**22**	**23**	**24**	**25**
jpn	62.6	82.2	90.2	87.9	53.9
kor	60.2	75.5	90.9	86.9	32.8
aus	59.8	75.4	90.7	88.9	36.0
usa	54.3	62.3	83.2	81.6	22.6

	M161Q01	M179Q01T	M192Q01T	M266Q01T	M273Q01T
Item No.	**26**	**27**	**28**	**29**	**30**
jpn	72.4	24.9	59.0	42.6	68.3
kor	63.1	28.0	52.0	35.9	57.3
aus	60.3	37.5	46.0	24.4	56.4
usa	45.5	30.0	28.0	12.8	47.6

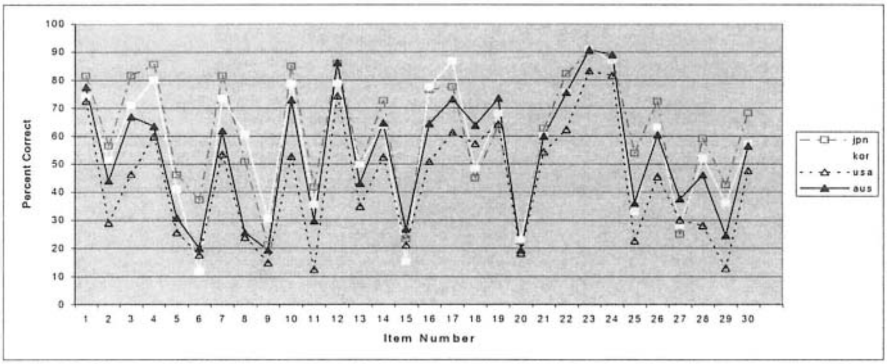


Figure 2-6-4. Plot of percentages correct by item and by country

5. DIFFERENTIAL ITEM FUNCTIONING

In this section, we examine the items that exhibit DIF. Further, assuming that there are differences between East and West traditions in mathematics, we test the hypothesis that when there is DIF, Japan and Korea are likely to be performing similarly as a group, and Australia and United States are performing similarly in another group. That is, the DIF is between these two groups of countries, rather than between the four individual countries.

To show DIF, we need to compare item difficulties across countries after making adjustments for the ability of students in those countries. That is, since we know that Japanese students are generally more proficient in mathematics than students in Australia, we expect that the percentages correct for Japan will be higher than those for Australia. Such differences in percentages correct are not an indication of DIF. However, after adjusting for ability, that is, comparing item difficulties given a student's ability, any observed differences in percentages correct would indicate the presence of DIF. We say that there is DIF when a student with the same ability in Japan would find an item more difficult, or easier, than a student of the same ability in Australia.

In addition, percentages correct, as a metric for measuring item difficulty, are known to have problems, mainly because of the bounded nature (between 0 and 1) of these measures. Therefore, for our analysis, we use item difficulty estimates obtained from calibrations of PISA data through item response modelling (IRM)[1] to examine DIF.

IRT analysis was carried out for each country separately. That is, the items were calibrated country by country. Table 3 gives the IRT item difficulty estimates by country. These item difficulty estimates are obtained after making adjustments for the ability level of students in a country. A high (more positive) value of item difficulty estimate indicates that the item is

Table 2-6-3. IRT item difficulty estimates (logits) by country

Item Code	**Item No.**	**Japan**	**Korea**	**Australia**	**United States**	**Mean**
M033Q01	**1**	-1.197	-0.925	-1.456	-1.598	-1.294
M034Q01T	**2**	0.315	0.281	0.595	0.852	0.511
M037Q01T	**3**	-1.399	-0.839	-0.867	-0.446	-0.888
M037Q02T	**4**	-1.700	-1.514	-0.543	-1.173	-1.233
M124Q01	**5**	0.602	0.804	1.013	0.802	0.805
M124Q03T	**6**	0.953	2.334	1.475	1.482	1.561
M136Q01T	**7**	-1.208	-0.909	-0.450	-0.669	-0.809
M136Q02T	**8**	0.536	-0.088	1.558	1.086	0.773
M136Q03T	**9**	2.015	1.201	1.668	1.355	1.560
M144Q01T	**10**	-1.707	-1.332	-1.137	-0.854	-1.258
M144Q02T	**11**	0.832	1.021	1.024	1.914	1.198
M144Q03	**12**	-1.865	-1.389	-2.267	-2.046	-1.892
M144Q04T	**13**	0.417	0.272	0.332	0.27	0.323
M145Q01T	**14**	-0.301	-0.244	-0.365	-0.271	-0.295
M148Q02T	**15**	1.915	2.256	1.441	1.297	1.727
M150Q01	**16**	-0.694	-0.936	-0.449	-0.321	-0.600
M150Q02T	**17**	-0.384	-1.016	-0.878	-0.896	-0.794
M150Q03T	**18**	0.980	0.617	-0.349	-0.620	0.157
M155Q02T	**19**	0.114	-0.423	-0.857	-1.015	-0.545
M155Q03T	**20**	1.829	1.616	1.938	1.528	1.728
M155Q04T	**21**	0.121	-0.172	-0.106	-0.424	-0.145
M159Q01	**22**	-0.919	-0.874	-1.017	-0.774	-0.896
M159Q02	**23**	-1.786	-2.12	-2.247	-2.205	-2.090
M159Q03	**24**	-1.479	-1.641	-2.098	-2.068	-1.822
M159Q05	**25**	0.617	1.455	1.189	1.456	1.179
M161Q01	**26**	-0.189	-0.100	-0.011	-0.002	-0.076
M179Q01T	**27**	2.167	1.520	1.086	1.020	1.448
M192Q01T	**28**	0.257	0.284	0.509	1.016	0.517
M266Q01T	**29**	1.237	1.279	1.869	2.160	1.636
M273Q01T	**30**	-0.081	0.179	0.220	-0.015	0.076

[1] Item response theory (IRT) was used in calibrating item difficulties in PISA. In particular, the generalised Rasch model (Wu, Adams & Wilson, 1997) was applied to item response data to estimate item difficulty parameters.

difficult, whilst a low (more negative) value indicates that the item is easy, for a person with some fixed ability. The item difficulty estimates are essentially a non-linear, monotonic transformation of the percentage correct, but adjusted for the ability level. The unit of the item difficulty estimate is 'logit', short for 'log of the odds'. If there is no DIF on an item, then we would expect the item difficulty estimates to be the same, within measurement errors, across the four countries. When the differences between the item difficulty estimates across the four countries are so great that they could not be explained simply by measurement errors, we would then conclude that DIF exists for this item. For example, for item 1, a student in Japan would find the item easier than a student with the same ability in Korea, but a student in Australia or the United States would be more likely than a Japanese student of the same ability to be successful.

Without formally carrying out statistical significance tests for DIF, we examine graphically the deviation of item difficulty estimates for each country from the mean item difficulty of the four countries for each item. Figure 5 shows the results. For example, the four points (-0.47, -0.28, 0.06, 0.69) plotted for item 4 shows that for Japan, the item difficulty is 0.47 logits lower than the average item difficulty for this item, after adjusting for ability level. For Korea, the item difficulty is 0.28 lower than the average. For the United States, the item difficulty is very close to the average item difficulty. But for Australia, students find this item relatively difficult as compared to students of similar ability in other countries, and one needs to add 0.69 logits to the average item difficulty to obtain the item difficulty estimate for Australia.

A wide range of points plotted for an item in Figure 5 shows that the item is not functioning in the same way in all four countries, while a clustering of the four points for an item indicates the item is functioning in the same way in all four countries.

When an item shows DIF in Figure 5, it is interesting to observe that the Eastern countries (Japan and Korea) and Western Countries (the United States and Australia) tend to group together. We use square symbols to indicate Eastern countries, and triangle symbols to indicate Western countries. For item 1, we see that the square symbols are on one side, and the triangle symbols are on the other side. We observe this pattern for many of the 30 items. For example, Eastern countries find items 2, 4, 7, 8, 10, 16, 28, 29 easier. Western countries find items 1, 12, 15, 18, 19, 23, 24, 27 easier. This is an indication that when there are deviations from the mean item difficulty, Japan and Korea tend to have the same kind of deviation, while the United States and Australia tend to have similar deviation as well.

Two questions come to mind regarding these observations: (1) Is the clustering of Eastern countries and Western countries happening by chance?

(2) When the clustering happens on opposite sides, that is, when Eastern countries find an item easier than Western countries, and when Western countries find an item easier than Eastern countries, can we identify item characteristics relating to these two directions of deviation?

To answer (1), we give an approximate assessment of the chance of observing 16 items out of 30 items showing clustering of Eastern countries and Western countries. Assuming that the clustering of four countries can happen in any order by chance, that is, it is equally likely to observe [(1,2), (3,4)], [(1,3), (2,4)], [(1,4), (2,3)], the chance of observing the clustering of Eastern countries and Western countries is one in three. From a binomial distribution with p=0.333 and n=30 items, the probability of observing 16 items or more with Eastern and Western clusters is 0.019. That is, there is only a 2% chance of observing 16 or more items with the East and West clustering. This is a small probability, so we conclude that the observed clustering of East and West countries is not likely to happen by chance.

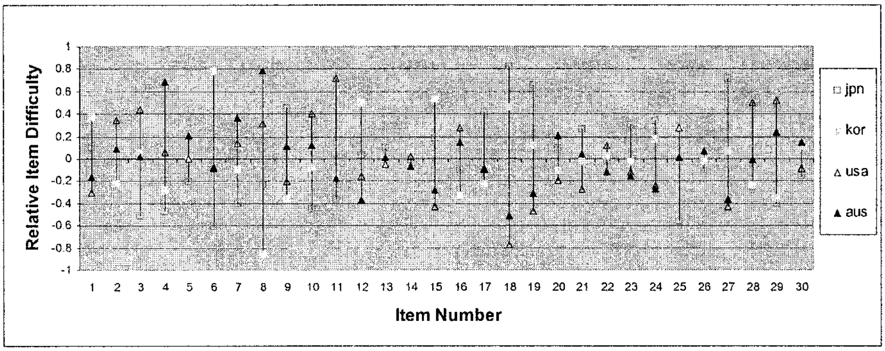


Figure 2-6-5. Deviations of item difficulty estimates from mean item difficulty

6. CLUSTER ANALYSIS

Another method to evaluate the "distances" between the four countries in terms of the item difficulty estimates is to carry out a cluster analysis. The data set for the cluster analysis consists of a matrix of 4 cases (corresponding to the 4 countries) and 30 variables (30 item difficulty estimates). A hierarchical cluster analysis is carried out to cluster cases (countries). A Dendrogram is produced as shown in Figure 6.

```
* * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * *


 Dendrogram using Average Linkage (Between Groups)

                       Rescaled Distance Cluster Combine

   C A S E    0         5        10        15        20        25
  Label  Num  +---------+---------+---------+---------+---------+

  aus      1
⇩×⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  usa      4   ⇩⇗                                              ⇔
  jpn      2
⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩×⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
  kor      3   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
```

Figure 2-6-6. Dendrogram from cluster analysis with 4 countries

Figure 6 shows that Australia and the United States are very close in their patterns of item difficulty estimates, with a distance of about 1 unit between the two countries, so these two countries form cluster 1. Next closest is Japan and Korea, at a distance of about 11 units, and these two countries form cluster 2. At a distance of around 25 units, the four countries form one large cluster. In summary, the cluster analysis identifies two clusters, with West countries in one cluster, and East countries in another cluster. Furthermore, the distance between the two West countries is much closer than the distance between the two East countries. But the two clusters are at least as far apart as the distances between countries within each cluster.

This result provides encouraging support for the hypothesis that Eastern and Western countries have consistent differences in their patterns of item facilities. Furthermore, these differences are evident in the PISA assessment and the differences are in some sense measurable. In view of the similarities between eastern European countries and Japan and Korea, as shown earlier in the comparison between PISA and TIMSS, we carried out a further cluster analysis with eight countries: Australia, Germany, Great Britain, Hungary, Japan, Korea, Russia, and the United States. The results of the cluster analysis are shown in Figure 7.

```
* * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * *

 Dendrogram using Average Linkage (Between Groups)

                         Rescaled Distance Cluster Combine

   C A S E     0         5        10        15        20        25
  Label  Num  +---------+---------+---------+---------+---------+

  aus       1   ⇩×⇩⇩⇩⇘
  gbr       7   ⇩⇗   □⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  usa       8   ⇩⇩⇩⇩⇩⇗            □⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  ger       2   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
□⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  hun       3   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
⇔
  jpn       4
⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩×⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘   ⇔
  kor       5   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
□⇩⇩⇩⇗
  rus       6
⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇗
```

Figure 2-6-7. Dendrogram from cluster analysis with 8 countries

Figure 7 shows the "distances" between the eight countries. The closest distance is between Australia and Great Britain. The United States joins this group next, followed by Germany, followed by Hungary. The other group starts with Japan and Korea in the cluster. They are joined by Russia before joining the other group to form one big cluster. This cluster analysis again shows that the grouping of Eastern and Western traditions is evident.

7. LINKING ITEM CHARACTERISTICS TO CULTURAL DIFFERENCES

Having established that there are indeed differences between Eastern and Western countries, we return to the question of whether we can identify item characteristics that can be linked to these differences. Table 4 is a table of items which Eastern countries find easier, and Table 5 shows a table of items which Western countries find easier.

The classifications of the items are mostly self-explanatory. We give only a few remarks. In the column headed "Big Idea (class)", the number in the brackets is the competency class number. There are three competency classes in PISA (OECD, 1999). These are:

Class 1: reproduction, definitions, and computations.
Class 2: connections and integration for problem solving.
Class 3: mathematical thinking, generalisation and insight.

Table 2-6-4. Items Eastern countries find easier

Item No.	**Item Code**	**Item Name**	**Item format**	**Response type (process)**	**Big Idea (class)**	**Mathematics Strand**	**Formal Mathematics**
2	M034Q01T	Not released	Closed Constructed Response	Numeric answer (counting)	Space and Shape (2)	Geometry	No
4	M037Q02T	Farms Q2	Closed Constructed Response	Numeric answer (geometric property)	Space and Shape (2)	Measurement	Yes
7	M136Q01T	Apples Q1	Closed Constructed Response	Numeric answer (pattern)	Growth and Change (2)	Algebra	Some
8	M136Q02T	Apples Q2	Closed Constructed Response	Numeric answer (equation)	Growth and Change (2)	Algebra	Yes
10	M144Q01T	Not released	Closed Constructed Response	Numeric answer (counting)	Space and Shape (1)	Geometry	No
16	M150Q01	Not released	Closed Constructed Response	Numeric answer (read graph)	Growth and Change (1)	Number	Some
28	M192Q01T	Not released	Multiple Choice	Match function	Growth and Change (2)	Measurement	Yes
29	M266Q01T	Not released	Multiple Choice	Assess property of shapes	Space and Shape (2)	Measurement	Yes

In the column headed "Formal Mathematics", we asked four experts to make judgments on whether an item contains mostly formal, curriculum-based content, or non-curriculum mathematics that nevertheless calls for sense making of real-world problems using mathematics. The judgments of the experts are averaged and summarised as *Yes*, *No* or *Some*. This exercise was not carried out with a stringent experimental design and control. It was done merely to seek some indications of item content. A wider consultation of this kind is necessary to have a fuller, and more accurate, evaluation of the processes involved in the items.

Table 2-6-5. Items Western countries find easier

Item No.	Item Code	Item Name	Item format	Response type (process)	Big Idea (class)	Mathematics Strand	Formal Mathematics
1	M033Q01	Not released	Multiple Choice	Spatial orientation	Space and Shape (1)	Geometry	No
12	M144Q03	Not released	Multiple Choice	Numeric Answer (counting)	Space and Shape (2)	Geometry	No
15	M148Q02T	Continent Area	Closed Constructed Response	Numeric Answer (estimation)	Space and Shape (2)	Measurement	Some
18	M150Q03T	Not released	Open Constructed Response	Verbal explanation of graph	Growth and Change (2)	Statistics	Some
19	M155Q02T	Not released	Closed Constructed Response	Numeric Answer (read unconventional graph)	Growth and Change (2)	Statistics	Some
23	M159Q02	Racing Car	Multiple Choice	Interpret graph	Growth and Change (1)	Functions	Some
24	M159Q03	Racing Car	Multiple Choice	Interpret graph	Growth and Change (1)	Functions	Some
27	M179Q01T	Not released	Open Constructed Response	Verbal explanation of graph	Growth and Change (2)	Functions	Some

What conclusions can we draw from Table 4 and Table 5? First, we note that Western countries are likely to perform better when the item content involves less formal mathematics. Second, Eastern countries perform well when an item involves numeric computation related to curriculum-based content, but they do not perform as well when an item calls for verbal explanations or interpretations of graphs. So the response type appears to have an impact on the performance between Eastern and Western countries. The item format (multiple choice or constructed) does not appear to make any difference in the relative performance between Eastern and Western countries; neither do Big Ideas nor Competency Classes. There may be a suggestion that Eastern countries do not perform as well in Statistics.

We give two examples to illustrate the key distinctions between the performance of Eastern and Western countries. Two items show large differences between Eastern and Western countries (see Figure 5): item 8 and item 15. These two items are given in the Appendix.

Item 8 (Apples Q2) requires students to form an equation and solve it, although students can use trial-and-error method as well. It is clear that both Japan and Korea performed extremely well on this item as compared to other items. This item calls for the use of formal mathematics learned in schools, including the use of symbolic representations of quantities.

Item 15 (Continent Area Q2) asks students to make an estimation of an irregular area. Many methods can be used. There is no single correct answer. Students are open to innovative ideas. They can use estimation methods learned in the classroom, or draw on their own experience in real-life to solve this problem. While they do need to understand the concept of area and scale units, the estimation method is completely open to their own creative resourcefulness.

These two examples highlight the item characteristics that make a difference to the performance of Eastern and Western countries. There are, however, many factors that have an impact on students' performance. Unfortunately, with the embargo on some of the items, it is difficult to illustrate these factors fully.

8. CONCLUSIONS

This paper demonstrates that there is indeed an Eastern tradition and a Western tradition in mathematics education. Further, these traditions are reflected in international comparative studies, and some characteristics of these traditions can be identified. What are the implications of these findings? We return to a question raised earlier about the validity of international studies. Clearly, having found the distinguishing features between Eastern and Western countries in their performance in mathematics, one can manipulate the content of an assessment to change the rankings of countries, particularly in relation to Eastern and Western cultures. PISA 2000 and TIMSS 1999 results suggest that differences in the balance of test material may result in a re-ordering of country performances. There are two implications of this observation. Firstly, the interpretation of international study results must be made in the light of the construct that is being tested. The term "Mathematics" has many different meanings to individuals, education specialists and policy makers. It is only meaningful when we report a country's relative standing in mathematics achievement when we clearly articulate what is being assessed. Secondly, in constructing any assessment, one must be careful about the nature of the items included and about the balance of the different kinds of items, as these can have a profound impact on the results. Mathematics educators must take an active part in deciding, reasoning and debating the kinds of mathematics competencies valued by the society in the 21st century. The revolution in information and communication technology has changed the world, and no doubt will continue to change the demands of skills and competencies in the workplace and in the home. Mathematics educators need to continue to adjust their goals to meet the demands of the changing world. What is the relative importance of being able to carry out formal mathematics procedures, or being able to communicate results to others, or being able to make sense of mathematical problems? These are questions we need to find answers to, before we can improve mathematics teaching and learning in schools.

Finally, this paper demonstrates that comparative studies can help us identify each country's strengths and weaknesses. Without international collaboration, we will not be able to make significant progress in making changes to educational practices. We also hope that the methodological approaches described in this paper, together with our findings, will stimulate further research in the area of international comparison.

REFERENCES

Adams, R.J. and Wu, M.L., 2002, *PISA 2000 Technical Report*. OECD. Paris.

Mullis, I.V.S., Martin, O.M., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., et al, 2000, *TIMSS 1999. International Mathematics Report*. International Study Centre, Boston College, Chestnut Hill.

Mullis, I.V.S., Martin, O.M., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., et al, 2001, *TIMSS Assessment Frameworks and Specifications 2003*. Boston College. Chestnut Hill.

OECD, 1999, *Measuring Student Knowledge and Skills: A New Framework for Assessment*. OECD Publications, Paris.

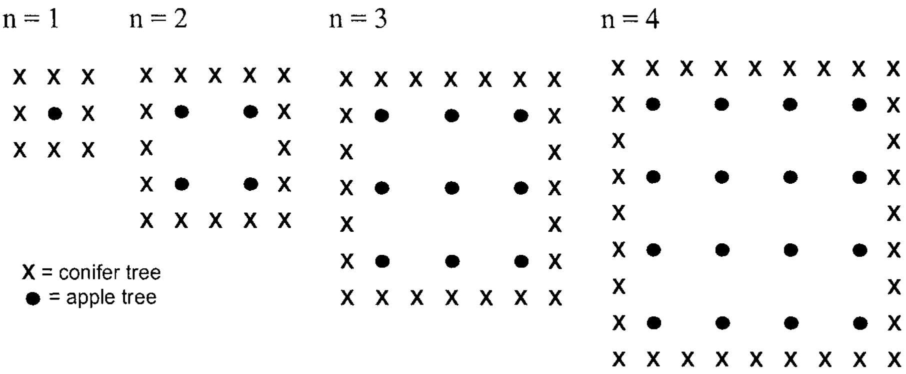
OECD, 2001. Knowledge and Skills for Life – First Results from PISA 2000. OECD Publications, Paris.

Wu, M.L., Adams, R.J., and Wilson, M.R., 1997, *ConQuest: Multi-Aspect Test Software.*, [computer program] Australian Council for Educational Research, Camberwell.

APPENDIX: SELECTED RELEASED PISA 2000 ITEMS

A farmer plants apple trees in a square pattern. In order to protect the apple trees against the wind he plants conifer trees all around the orchard.

Here you see a diagram of this situation where you can see the pattern of apple trees and conifer trees for any number (n) of rows of apple trees:



APPLES QUESTION 1 (ITEM 7, M136Q01)

Complete the table:

n	*Number of apple trees*	*Number of conifer trees*
1	1	8
2	4	
3		
4		
5		

APPLES QUESTION 2 (ITEM 8, M136Q02)

There are two formulae you can use to calculate the number of apple trees and the number of conifer trees for the pattern described above:

Number of apple trees = n^2

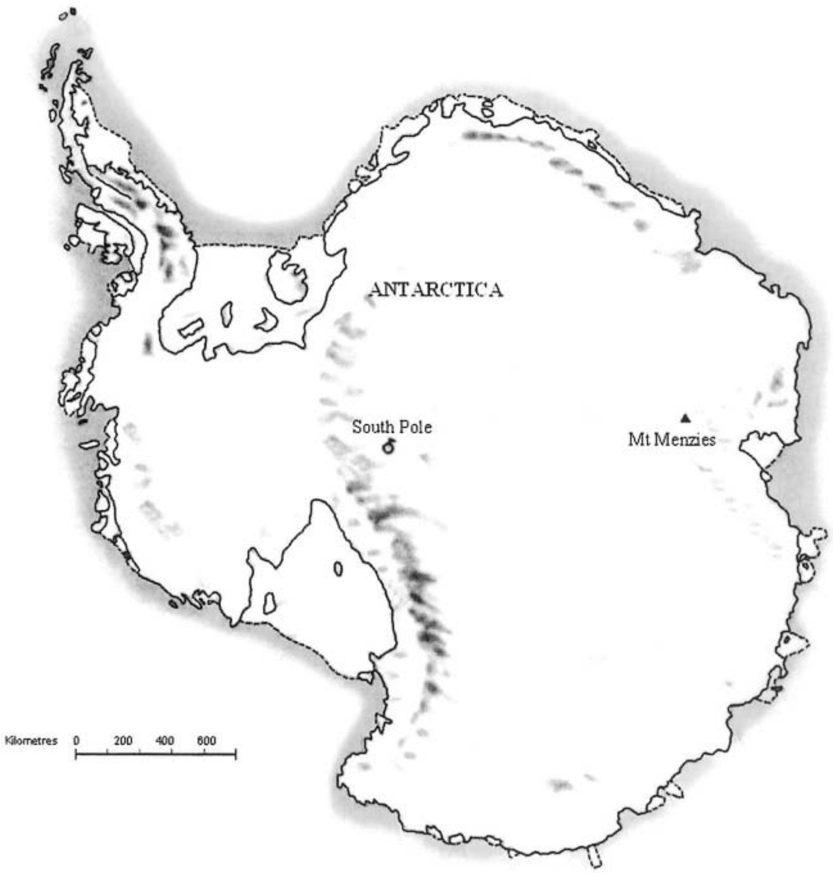
Number of conifer trees = $8n$

where n is the number of rows of apple trees.

There is a value of n for which the number of apple trees equals the number of conifer trees. Find the value of n and show your method of calculating this.

CONTINENT AREA (ITEM 15, M148Q02)

Below is a map of Antarctica



Estimate the area of Antarctica using the map scale.
Show your working out and explain how you made your estimate. (You can draw over the map if it helps you with your estimation)