Estimating Trends in National Performance from International Surveys, with a focus on PIRLS Results for England

By

Ian Schagen[1]

Liz Twist[2]

Simon Rutt[2]

[1]Ministry of Education, Wellington, New Zealand (formerly at NFER)

[2]National Foundation for Educational Research (NFER), UK

Abstract

No man ever steps in the same river twice, for it's not the same river and he's not the same man. (Heraclitus)

England's results for PIRLS 2006 showed a significant decline from 2001, in absolute scores as well as relative rankings. Closer examination of the performance on common items between the two years did not bear this out, so further analysis has been carried out. A recent paper by Gebhardt and Adams (2007) indicated that equating methodology affects trend estimates, with examples from PISA. We have followed up this work by investigating different approaches to estimating PIRLS trends for England, resulting in substantively different results from the same data, depending on the methodology used.

Keywords: trend analysis, item response theory, PIRLS.

Introduction

The results of the first PIRLS survey (PIRLS 01) were notable for England's high performance. Of the 35 participating countries, only one (Sweden) scored significantly higher.

The results of PIRLS 06, the second survey, were published in November 2007. In both absolute terms (scale score) and relative terms (in comparison with other participating countries) the international report shows that England's performance was poorer than in 2001.

Before looking in more detail at the outcomes of PIRLS 06, it is useful to provide some background to the debate about standards in England which exists quite independently of the international survey results.

1 England's national assessment system

England has an extensive national assessment system. At four points in their school careers, all students will undertake some form of external assessment (Table 1).

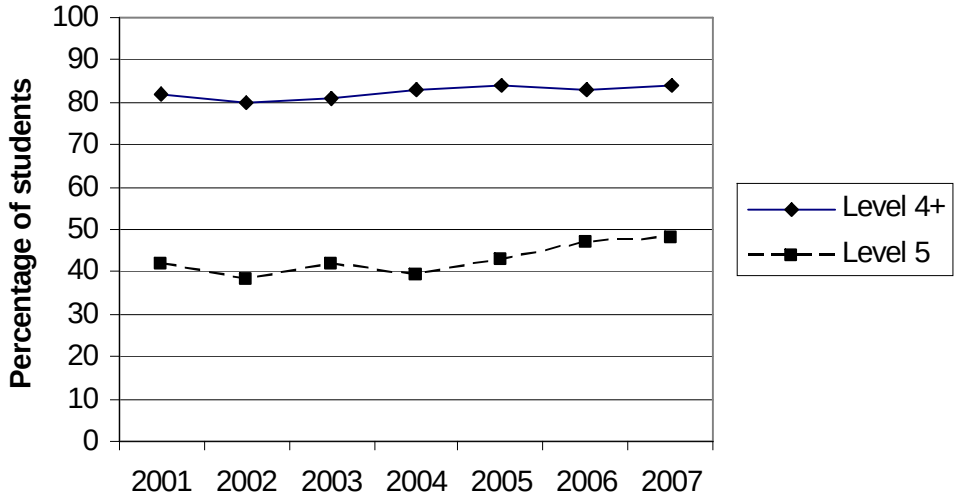
Tests at age 11 are of less significance for individual students than those at age 16, but are of considerable importance to the 17,000 primary schools, whose success is often judged on the basis of their ranking.

The tests for students aged 11 are on a four point scale. At this age, students are expected to reach level 4 and some reach level 5. There are national, regional and school targets for the proportion of students at these two levels. Figure 1 shows percentage of students reaching level 4 or above, and level 5, from 2001 to 2007. These students are in grade 5 whereas PIRLS students are in grade 4.

Table 1: National assessments in England

Age	Nature of assessment	Extent	Significance / importance of outcomes
7	Tests in reading, writing and mathematics	Whole cohort testing	Reported to parents as part of an overall 'teacher assessment'. Results collated nationally.
11	Tests in reading, writing, mathematics and science	Whole cohort testing	Reported to parents (alongside 'teacher assessments') and to secondary schools. Tables produced to enable comparison of one school with another. Schools have their own targets. Local authority tables. National targets.
14	Tests in reading, writing, mathematics and science	Whole cohort testing	Reported to parents (alongside 'teacher assessments'). Tables produced to enable comparison of one school with another. Schools have their own targets. Local authority tables. National targets.
16	Public examinations in many different subjects	Selected subjects incl. certain prescribed subjects	16 is school leaving age for some students. Can determine what future course of study student is eligible for. Tables produced to enable comparison of one school with another. Schools have their own targets. Local authority tables. National targets.

Figure 1: Percentage of students achieving level 4+ in national reading test at age 11



It is apparent from Figure 1 that the proportions of students reaching the national target at age 11 has remained broadly stable since 2002 (the year in which the PIRLS 01 students took this test) with possibly a slight increase over the five years to 2007. As there are quite large percentages of students around each score point at the thresholds, a movement of one or two percentage points a year is not very meaningful. There does seem to be a more stable trend when the proportions of students achieving the highest level are considered. This figure was 38% in 2002 and was 48% in 2007, having recorded a steady increase from 2004.

Whilst there is occasionally some skepticism about the national results, as representatives, formerly and currently, of the development agency for these tests, the authors are aware that great efforts are made to ensure standards over time are maintained, given that we are charged with producing completely new assessments each year.

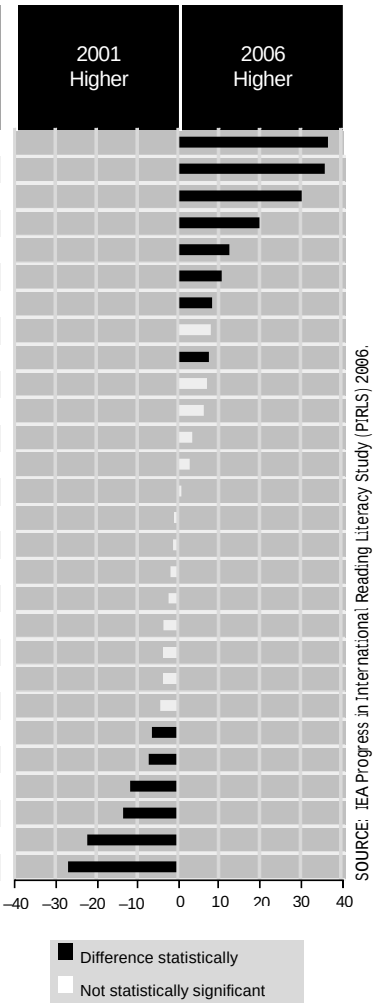
2 PIRLS 2001 and 2006 trend results

It was against this background that the results of PIRLS 06 were released and not unexpectedly attracted political, media and academic interest. As the researchers involved we were not surprised that many countries showed an increase in their overall scale score; what did surprise us somewhat was that England, and other specific countries, registered a significant fall.

Figure 2 shows the trend in terms of scale scores from 2001 to 2006 for countries participating in both surveys (Mullis, Martin, Kennedy, and Foy, 2007).

Figure 2: Trends in reading achievement, PIRLS 2001-2006

Countries	PIRLS 2006 Average Scale Score	PIRLS 2001 Average Scale Score	Difference Between 2001 and 2006 Scores
2a Russian Federation	565 (3.4)	528 (4.4)	37 (5.6)
Hong Kong SAR	564 (2.4)	528 (3.1)	36 (3.9)
Singapore	558 (2.9)	528 (5.2)	30 (5.9)
Slovenia	522 (2.1)	502 (2.0)	20 (2.9)
Slovak Republic	531 (2.8)	518 (2.8)	13 (4.0)
Italy	551 (2.9)	541 (2.4)	11 (3.8)
Germany	548 (2.2)	539 (1.9)	9 (2.9)
Moldova, Rep. of	500 (3.0)	492 (4.0)	8 (5.0)
Hungary	551 (3.0)	543 (2.2)	8 (3.7)
Iran, Islamic Rep. of	421 (3.1)	414 (4.2)	7 (5.2)
2a *Canada, Ontario*	554 (2.8)	548 (3.3)	6 (4.4)
2b Israel	512 (3.3)	509 (2.8)	4 (4.4)
New Zealand	532 (2.0)	529 (3.6)	3 (4.1)
Macedonia, Rep. of	442 (4.1)	442 (4.6)	1 (6.2)
† Scotland	527 (2.8)	528 (3.6)	-1 (4.6)
‡ Norway	498 (2.6)	499 (2.9)	-1 (3.9)
Iceland	511 (1.3)	512 (1.2)	-2 (1.8)
†2a United States	540 (3.5)	542 (3.8)	-2 (5.2)
2a Bulgaria	547 (4.4)	550 (3.8)	-3 (5.8)
France	522 (2.1)	525 (2.4)	-4 (3.1)
Latvia	541 (2.3)	545 (2.3)	-4 (3.3)
Canada, Quebec	533 (2.8)	537 (3.0)	-4 (4.1)
Lithuania	537 (1.6)	543 (2.6)	-6 (3.1)
† Netherlands	547 (1.5)	554 (2.5)	-7 (2.9)
Sweden	549 (2.3)	561 (2.2)	-12 (3.2)
England	539 (2.6)	553 (3.4)	-13 (4.3)
Romania	489 (5.0)	512 (4.6)	-22 (6.8)
Morocco	323 (5.9)	350 (9.6)	-27 (11.3)



† Met guidelines for sample participation rates only after replacement schools were included (see Exhibit A.7).
‡ Nearly satisfying guidelines for sample participation rates after replacement schools were included (see Exhibit A.7).
2a National Defined Population covers less than 95% of National Desired Population (see Exhibit A.4).
2b National Defined Population covers less than 80% of National Desired Population (see Exhibit A.4).
( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Trend Note: The primary education systems of the Russian Federation and Slovenia underwent structural changes. Data for Canada, Ontario include only public schools.

It is apparent from Figure 2 that the fall in England's performance since the first survey is greater than for all other countries with the exception of Romania and Morocco. What is equally striking is that the three top performing countries in 2001 (Sweden, the Netherlands and England) all recorded a fall in their overall scale scores.

Whilst the presentation of the data in Figure 2 draws attention to the relative change across countries, within England, however, the interest was on the absolute fall from 2001 to 2006. Monseur, Sibberns and Hastedt (2007) recognise this focus on how within-country differences evolve over time and suggest that "the political importance of Trends in achievement should not be underestimated".

In addition to producing an overall scale score, PIRLS also scales performance on the two reading purposes identified in the reading framework (Mullis, Kennedy, Martin, and Sainsbury, 2006). Table 2 summarises performance of the top 10 performing countries in 2001, and shows the scale scores achieved in 2001 and 2006 on the main scale and the two reading purposes subscales.

Table 2: Difference in scale scores, 2001 - 2006, for the ten highest achieving countries in 2001

	Overall scale score			**Literary reading scale score**			**Informational reading scale score**		
	2006	2001	difference	2006	2001	difference	2006	2001	difference
Sweden	549	561	-12	546	559	-13	549	559	-10
Netherlands	547	554	-7	545	552	-8	548	553	-5
England	539	553	-13	539	559	-20	537	546	-9
Bulgaria	547	550	-3	542	550	-7	550	551	-1
Latvia	541	545	-4	539	537	+2	540	547	-7
Lithuania	537	543	-6	542	546	-4	530	540	-10
Hungary	551	543	+8	557	548	+8	541	537	+4
United States	540	542	-2	541	550	-10	537	533	+4
Italy	551	541	+11	551	543	+8	549	536	+13
Germany	548	539	+9	549	537	+12	544	538	+6

Countries ordered by 2001 overall scale score

Among the subset of high achieving countries in 2001, overall scale scores rose by up to 11 points (Italy) and fell by up to 13 points (England). On the literary scale, the greatest rise was of 12 points (Germany) and the greatest fall was of 20 points (England). On the informational reading scale, the greatest rise was again Italy with a rise of 13 points, and the greatest fall was of ten points (Sweden and Lithuania).

Whilst Table 2 takes the 10 highest performing countries in 2001 and shows how the overall scale score of seven fell, including the scale scores of the six highest achieving countries, it does not include the three highest achieving countries in the 2006 survey as

these were not among the highest achievers in 2001. These countries are shown in Table 3.

Table 3: Difference in scale scores, 2001 - 2006, for the three highest achieving countries in 2006

	Overall scale score			Literary reading scale score			Informational reading scale score		
	2006	2001	difference	2006	2001	difference	2006	2001	difference
Russian Fed.	565	528	+37	561	523	+38	564	531	+32
Hong Kong	564	528	+36	557	518	+39	568	537	+31
Singapore	558	528	+30	552	528	+23	563	527	+36

Further data illustrating the trend in performance in 2001 and 2006 is included in the international report (Mullis, Martin, Kennedy and Foy, 2007) including performance in the two surveys of boys and girls, and performance on the two scales illustrating the four reading processes assessed in PIRLS. These are not reviewed here.

The rest of this paper investigates further the data detailing performance of the three highest achieving countries in PIRLS 01: Sweden, the Netherlands and England.

3 Investigation into performance on the linking items in 2001 and 2006

When the overall results for 2001 are compared with the results for 2006 for England, we have the picture as shown in outline in Table 4 below.

Table 4: England's results in 2001 and 2006 (international database)

Scale	2001		2006		Difference (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Literary	558.0	88.5	539.8	84.8	-18.2
Information	546.7	77.9	537.7	78.8	-9.0
Overall	552.9	82.7	539.5	83.3	-13.4

(Results weighted to allow for sample non-response)

Only four assessment blocks were common to the 2001 and 2006 tests (the 'linking blocks'), and therefore any changes must be based on linking via these blocks. Total item scores on these blocks for England were compared, in order to see if the pattern of performance on these items corresponded to the results presented above. These values are

shown in Table 5 below. Detailed item-level comparisons of the data for all three countries were conducted but are not included here.

Table 5: England's results in 2001 and 2006 (common assessment blocks)

Block (purpose)	2001		2006		Difference in mean (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Lit. C	10.21	4.35	10.50	4.30	+0.29
Lit. F	11.59	3.63	11.07	3.62	-0.52
Inf. A	11.90	3.05	10.89	3.30	-1.01
Inf. L	9.27	3.96	8.91	3.97	-0.36

(Results weighted to allow for sample non-response)

These simple comparisons suggest a different picture from that shown by the scaled measures. Changes in scores on the literary items seem to be in approximate balance in terms of common items, whereas the biggest change in scaled score is for the literary outcomes. The opposite is true for information – a relatively large decline (especially in Block A) matched by a lower decline in scaled scores.

The same analysis was then undertaken for the data from Sweden and from the Netherlands and the equivalent tables are shown below (Tables 6-9).

Table 6: Sweden's results in 2001 and 2006 (international database)

Scale	2001		2006		Difference (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Literary	560.4	59.4	545.9	57.6	-14.5
Information	558.7	63.1	549.2	62.3	-9.5
Overall	561.0	61.5	549.3	59.8	-11.7

(Results weighted to allow for sample non-response)

Table 7: Sweden's results in 2001 and 2006 (common item blocks)

Block (purpose)	2001		2006		Difference in mean (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Lit. C	11.26	3.36	10.94	3.42	-0.32
Lit. F	12.14	2.99	11.70	2.99	-0.44
Inf. A	12.30	2.73	12.11	2.78	-0.19
Inf. L	8.80	3.27	9.05	3.12	0.25

(Results weighted to allow for sample non-response)

For Sweden, these simple comparisons suggest a similar picture to that shown by the scaled measures. The performance in 2006 for both literary blocks shows a decline in performance when compared to 2001. For information there is a decline for Block A combined with a slightly larger rise for Block L.

The analysis of the data from The Netherlands is shown in the following two tables.

Table 8: The Netherlands' results in 2001 and 2006 (international database)

Scale	2001		2006		Difference (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Literary	551.6	53.0	544.8	51.5	-6.8
Information	553.0	53.4	547.7	45.2	-5.3
Overall	554.2	53.2	547.2	49.1	-7.0

(Results weighted to allow for sample non-response)

Note that one item in Block L (R011L06C) was completely missing for all cases in 2001. The comparisons for the Netherlands are therefore made omitting this item from Block L in both years, and all further analyses also omit this item.

Table 9: The Netherlands' results in 2001 and 2006 (common item blocks)

Block (purpose)	2001		2006		Difference in mean (2001-2006)
	Mean	Standard deviation	Mean	Standard deviation	
Lit. C	10.65	3.49	11.09	3.40	+0.44
Lit. F	11.26	2.86	11.55	2.75	+0.29
Inf. A	12.30	2.43	12.01	2.36	-0.29
Inf. L	8.89	2.85	9.23	2.74	+0.34

(Results weighted to allow for sample non-response)

As with the data from England, these simple comparisons create a different picture from that shown by the scaled measures. Changes in scores on the literary items seem to indicate an overall rise in terms of common items, whereas the biggest negative change in scaled score is for the literary outcomes. For information there is a decline for Block A matched by a rise for Block L, and a lower overall decline in scaled scores.

To summarise the data in Tables 4-9:

- On the basis of the analysis in the international report, the performance of all three countries fell in 2006 on the overall scale and on both subscales.

- In the IEA analysis, all three countries showed a greater fall on the literary subscale than on the informational subscale. For England, the literary scale fell 20 scale points, compared to a fall on the information scale of 9 scale points; for Sweden it was 13 lower (literary) and 10 lower (information); for the Netherlands it was 10 lower (literary) and 5 lower (information).

- When a simple comparison is made of the outcomes on just the linking blocks, the raw scores from England fell on three blocks and rose on one (block C, literary); the raw scores from Sweden also fell on three and rose on one (block L, informational); in respect of the Netherlands, the raw scores rose on three and fell on one (block A, informational).

On the surface these were puzzling results which required further scrutiny. Two investigations were carried out. Firstly, the linking methodology was scrutinised to see if the method adopted could have an impact on the outcomes. This is summarised briefly below. Secondly, items in which facilities appeared to show the greatest change between 2001 and 2006 in the three selected countries were studied. This element is not reported here.

4 Investigating the linking methodology

The first stage of the investigation is to attempt to link each country's results in 2001 and 2006 using a two-parameter IRT model via the common items.

4.1 IRT equating results – individual countries – joint calibration

IRT equating results (England only) – joint calibration with data from PIRLS 2001 and PIRLS 2006

Data from all items and all students in England in 2001 and 2006 was combined into a single file, including the common items and those specific to each year (174 in total). A total of 7369 cases were included in the analysis. A two-parameter IRT model was fitted to the dataset using appropriate software[1], and the fit of the model to the item data was seen to be good. From this model student ability values were estimated, and these were rescaled so that the 2001 results (weighted) gave the same mean and standard deviation as those produced for the international report.

A similar process was undertaken for literary and information items separately (89 and 85 items), in order to replicate the performance measures for these aspects of reading. Results were scaled in a similar way to match the 2001 international report figures, and the full set of comparisons is shown in Table 10 below.

Table 10: England's results in 2001 and 2006 (linked via 2-parameter model for England only)

Scale	2001		2006		Difference in scale scores
	Mean	Standard deviation	Mean	Standard deviation	
Literary	558.0	88.5	559.4	90.1	+1.4
Information	546.7	77.9	540.4	78.9	-6.3
Overall	552.9	82.7	548.1	83.9	-4.8

(Results weighted to allow for sample non-response)

A simple multilevel analysis of these scaled scores showed that none of the above differences are statistically significant, once we take into account the clustering of students within schools and the corresponding design effect.

[1] Initially using the program PARSCALE; when this failed to converge in some circumstances models were fitted using the program MPLUS. Checks were made to ensure both programs gave identical results.

These results correspond more closely to the simple analysis presented in Table 5, but are strikingly at variance with Table 4.

The same analysis was then conducted using the data from Sweden and from the Netherlands.

IRT equating results (Sweden only) – joint calibration

Data from all items (174) and all students (10,438) in Sweden in 2001 and 2006 was combined into a single file. A two-parameter IRT model was fitted to the dataset and the fit of the model to the item data was seen to be good. As with England's data, model student ability values were estimated, and these were rescaled so that the 2001 results (weighted) gave the same mean and standard deviation as those produced by the IEA. A similar process was undertaken for literary and information items separately (89 and 85 items). Results were scaled in a similar way to match the 2001 international report figures, and the full set of comparisons is shown in Table 11.

Table 11: Sweden's results in 2001 and 2006 (linked via 2-parameter model for Sweden only)

Scale	2001		2006		Difference in scale scores
	Mean	Standard deviation	Mean	Standard deviation	
Literary	560.4	59.4	557.9	59.5	-2.5
Information	558.7	63.1	559.9	62.2	1.2
Overall	561.0	61.5	558.9	61.0	-2.1

(Results weighted to allow for sample non-response)

A simple multilevel analysis of these scaled scores showed that for the Information scale there is no statistical difference between the years once the clustering of students within schools and the corresponding design effect are taken into account. For the Literary scale and the overall scaled score there is a significant difference between 2006 and 2001, with results significantly lower in 2006, although the level of significance could not be described as very strong. The outcomes from this analysis more closely resemble the outcomes shown in Table 7 (common items) than those in Table 6 (IEA results).

The same analysis was conducted with data from the Netherlands.

IRT equating results (the Netherlands only) – joint calibration

Data from all items (173 in total, omitting R011L06C) and all students (8,268) in the Netherlands in 2001 and 2006 was combined into a single file, including the common items and those specific to each year. The analysis was conducted as described above.

A similar process was undertaken for literary and information items separately (89 and 84 items), in order to replicate the performance measures for these aspects of reading. Results were scaled in a similar way to match the 2001 international report figures, and the full set of comparisons is shown in Table 12 below.

Table 12: The Netherlands' results in 2001 and 2006 (linked via 2-parameter model for the Netherlands only)

Scale	2001		2006		Difference in scale scores
	Mean	Standard deviation	Mean	Standard deviation	
Literary	551.6	53.0	554.5	53.0	+2.9
Information	553.0	53.4	551.6	51.1	-1.4
Overall	554.2	53.2	555.2	51.5	+1.0

(Results weighted to allow for sample non-response)

A simple multilevel analysis of these scaled scores has shown that none of the above differences are statistically significant, once the clustering of students within schools and the corresponding design effect are taken into account. The outcomes from this analysis more closely resemble the outcomes shown in Table 9 (common items) than those in Table 8 (IEA results).

To summarise the outcomes of this part of the investigation:

When the data from both surveys is combined, and the equating is based on the data for each country separately, the results in all three cases correspond more closely to the simple analysis presented in Tables 5, 7 and 9 but are strikingly at variance with Tables 4, 6 and 8 respectively.

It is not immediately obvious why there should be this difference between the results obtained by linking a single country's 2001 and 2006 data from the results IEA obtain linking the years using all countries' data.

A possible explanation could relate to improved performance by a number of countries between 2001 and 2006, which could affect the item parameters and hence scores when linking using all countries.

4.2 IRT equating results – England only – calibration based on data from PIRLS 01

Because the above results gave a different picture from those reported in the international report, the calibration using England's data only was repeated using an alternative methodology. In principle, there are a number of ways in which such linking across years using common items can be carried out, including:

1. Fit a model for the early year's data only, and use the common item parameters from this model as fixed values for the model fitting the later year's data;

2. As (1), but fixing common item parameters based on the later year's data.

3. Form a single model spanning both years, with common item parameters estimated from both years' data (as above).

The IEA appears to have adopted the third method in the analysis of PIRLS 01 and PIRLS 06. As an alternative approach, and to investigate the sensitivity of the results to the calibration approach used, the data was analysed using method 1.

A two-parameter model was fitted to the 2001 data, and common item parameters in the model for the 2006 data were fixed to have the same values. Final ability values were again rescaled to have the same mean and variance in 2001 as for the IEA results. Table 13 shows the results of this calibration exercise with data from England.

Table 13: England's results in 2001 and 2006 (linked via 2-parameter model for England only, common items fixed for 2001)

	2001		2006		
Scale	**Mean**	**Standard deviation**	**Mean***	**Standard deviation***	**Difference***
Literary	558.0	88.5	558.1	88.0	0.0
Information	546.7	77.9	529.6	77.3	-17.2
Overall	552.9	82.7	543.1	84.0	-9.9

(Results weighted to allow for sample non-response)

*Preliminary data

Comparison of these results with those reported in Table 10, based on the same data, shows a clear difference related to the linking methodology. This is particularly the case for the information items, where the apparent decline has more than doubled. The change

in literary scores (approximately zero) is fairly consistent, but overall scores decline by roughly twice the amount shown in the previous analysis. This decline in overall scores is not quite statistically significant when design effects are taken into account, although the decline in information scores is significant.

4.3 Tentative conclusions and possible explanations

The exercise of linking 2001 and 2006 data for England, the Netherlands and Sweden (separately) via a two-parameter IRT model and common items has shown the following:

- For England, results overall on the linking literary blocks are very comparable between 2001 and 2006. Performance on the information blocks show a fall.
- For Sweden there appears to be little change on either the literary or the information blocks between 2001 and 2006.
- For the Netherlands, results overall on the linking literary and information blocks are very comparable in 2001 and 2006.
- At item level (detailed data not reported here), results on a few information items in England's data seem to have fallen between 2001 and 2006, leading to an overall slight reduction in information scores.
- Sweden's data is rather more variable, with quite a high proportion of items recording either an increase or a decrease in facility of five per cent or more. This applies to both literary texts and to one of the information texts.
- Data from the Netherlands shows that there were increased facilities on three of the four blocks. There were three items where facilities fell by five per cent or more, all on the Antarctica block.
- For England, none of the changes in overall or literary scores between 2001 and 2006 are significant when design effects are accounted for using multilevel modelling. However, when a linking method which calibrates against 2001 item parameters is used, there is an apparently significant decline in information scores.
- For Sweden, there is no significant difference between the informational scores in 2001 and 2006; however, when a linking method which calibrates against 2001 item parameters is used, there is an apparently significant decline in literary scores and in the overall score.
- For the Netherlands with this method, there is no significant difference between any of the scale scores in 2001 and 2006.

- Results are sensitive to the linking methodology.

Possible explanations for the discrepancies between the above findings and those produced in the PIRLS 06 international report include:

- Linking performance across years via common items with countries some of whose results have increased significantly may adversely affect the apparent scores of countries which did well in 2001.

- IEA results are based on 'plausible values' which take background factors into account, whereas the results presented here are based purely on test performance.

5 Gebhardt and Adams' paper (2007)

During the course of this investigation into the trend data in PIRLS 01 and PIRLS 06, a paper was published by Gerhardt and Adams who are closely involved with the PISA study.

The authors illustrate their paper with data from PISA 2000 and 2003 on reading and science, but their discussion and conclusions could apply to the estimation of trends from any international survey. They begin by considering what they refer to as the 'original' trend estimate for a given country. As with all trend estimates, it is crucially based on the link items which are common to the surveys in both years.

The first issue discussed is whether the item parameters (i.e. difficulties) for the link items should be the same for all countries (international estimates) or vary by country (national estimates). The former approach is used in PISA and all other international studies to produce trends, but the authors argue that this ignores item-by-country interactions which "are commonly observed in cross-national studies … and the magnitude of these interactions influences the validity of cross-country comparisons" (Gebhardt and Adams, 2007, p.307). In other words, if we estimate link item parameters for any individual country these may differ significantly from the international estimates, and this will affect the estimate of trend for that country.

The second issue they discuss is how the scales for the two studies are linked. Given the use of IRT methods in all such studies, it is important to have item difficulty values for these link items. As noted earlier, these can be estimated in three ways:

1. Based on data for the earlier study, and fixed for the later study;

2. Based on data for the later study, and fixed for the earlier study;

3. Based on combined data for both studies.

Method 2 above is apparently used for PISA; both methods 1 and 2 are described as 'linear transformation' in their Table 1, but in the paper Gebhardt and Adams argue for the use of method 3, which they describe as 'joint calibration'. It has the advantage that item parameters are estimated more precisely based on the full data, and the underlying student ability values (which are used to derive trends) are estimated directly.

5.1 Relationship to England's results for PIRLS

The first two issues raised by Gebhardt and Adams have been explored in the investigation and analyses described above in relation to the PIRLS data for England, the Netherlands and Sweden.

Every estimate of trends is crucially based on performance on the link items in both years, and in some ways the most direct evaluation of changes in national performance is derived from direct comparison of (weighted) mean scores on common item blocks. This was done in Tables 5, 7 and 9 above, and this seemed to show results which conflicted with the trend values produced internationally for all three countries. Further investigation included carrying out equating exercises for England alone (national item parameter estimation), using both joint calibration (method 3) and fixing link item parameters to their 2001 values (method 1). Although there were differences in the results obtained, in both cases there was a significant variance from the international results. In contrast to the latter, performance on the literary scale was stable in England from 2001 to 2006, while performance on the information scale fell somewhat.

These results bear out Gebhardt and Adams' conclusions that there are significant effects due to the choice of equating methodology, in particular related to the use of international rather than national parameter estimates.

5.2 Conclusions

There is a great deal of consensus between the work of Gebhardt and Adams on PISA data and the more limited investigation of PIRLS reported here. There is no one single equating methodology which both links countries and produces robust estimates of change over time. Within England, and possibly many other countries, the main focus of the trend measure is the change in the country's performance across two or more time points. It is not the extent of the change in performance in relation to the change in performance in other countries. It may be that to produce robust estimates of change over time in an informative way for individual countries we need to turn to a national parameter estimation system, as we have begun to explore for PIRLS in this paper. These analyses require complex models – we just need to ensure we are using the right model to answer the question we are posing.

The quote from Heraclitus is apposite – we are giving a different test to different students; only our conceptual model tells us that we are measuring the same thing and allows us to draw conclusions.

References

Foy, P., Galia, J. and Li, I. (2007). *Scaling the PIRLS 2006 Reading Assessment Data*. In Martin, M., Mullis, I. and Kennedy, A. (Eds.) PIRLS 2006 Technical Report. Chestnut Hill, MA: Boston College.

Gebhardt, E. and Adams, R.J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, 8, 3, 305-322.

Monseur, C., Sibberns, H. and Hastedt, D. (2007). Equating errors in international surveys in education. *Proceedings of the IRC-2006, Volume 2*. Amsterdam: IEA.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O. and Sainsbury, M. (2006). *PIRLS 2006 Reading Assessment Framework and Specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. and Foy, P. (2007). *PIRLS 2006 International Report*. Chestnut Hill, MA: Boston College.