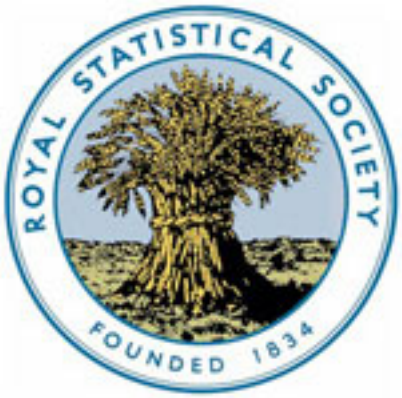

J. R. Statist. Soc. A (2007)
170, *Part* 3, *pp.* 623–646

International surveys of educational achievement: how robust are the findings?

Giorgina Brown,

Istituto Nazionale di Statistica, Rome, Italy

John Micklewright and Sylke V. Schnepf

University of Southampton, UK

and Robert Waldmann

University of Rome "Tor Vergata", Italy

[Received January 2005. Final revision September 2006]

Summary. International surveys of educational achievement and functional literacy are increasingly common. We consider two aspects of the robustness of their results. First, we compare results from four surveys: the Trends in International Maths and Science Study, the Programme for International Student Assessment, the Progress in International Reading Literacy Study and the International Adult Literacy Survey. This contrasts with the standard approach which is to analyse just one survey in isolation. Second, we investigate whether results are sensitive to the choice of item response model that is used by survey organizers to aggregate respondents' answers into a single score. In both cases we focus on countries' average scores, the within-country differences in scores and on the association between the two.

Keywords: Educational achievement; International Adult Literacy Survey; Programme for International Student Assessment; Progress in International Reading Literacy Study; Test scores; Trends in International Maths and Science Study

1. Introduction

Recent years have seen several international surveys of educational achievement of children and 'functional' literacy of adults: the 1994–1998 International Adult Literacy Survey (IALS), the 1995, 1999 and 2003 Trends in International Maths and Science Study (TIMSS), the 2000 and 2003 Programme for International Student Assessment (PISA) and the 2001 Progress in International Reading Literacy Study (PIRLS). Further survey rounds are planned. The existing data are already used widely by governments and international organizations and by researchers from various disciplines, e.g. the UK Government in Social Exclusion Unit (2001), the human poverty index 2 in United Nations Development Programme (2000) and, from disciplines outside education, Denny (2002) in social statistics, Wößmann (2003) in economics and Esping-Andersen (2004) in sociology.

One feature of all this activity is that the surveys are typically analysed in isolation from one another with no indication about whether new results confirm or contradict those from earlier surveys. But each survey has its merits and defects, and its own particular focus. The subjects

Address for correspondence: John Micklewright, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: jm4@soton.ac.uk

 0964–1998/07/170623

investigated, the age groups studied, the form of the tests and the survey response rates all vary. The results from different surveys therefore need to be compared. There have been valuable contributions to this endeavour, usually focused on a few countries, a pair of surveys and one subject (e.g. O'Leary *et al.* (2000), O'Leary (2001) and Prais (1997, 2003)). But to our knowledge no study has pulled together results for all subjects from all the surveys mentioned above for a large group of countries to compare key dimensions of the pattern of their results. Making such a comparison is the first contribution of this paper.

Comparing findings across surveys is one aspect of a search for robust results. Another is to explore the sensitivity of results to the choice of method for aggregating answers by each individual to a survey's questions into a single score. This aggregation is done by the surveys' organizers using item response models from the psychometric literature. In contrast with the more obvious issues listed above, such as subject or age group, most users of the achievement surveys are probably unaware that there is even an issue here of potential importance. The so-called 'scaling' methods of the item response models have been questioned by some commentators and alternative models have been applied to the data for selected countries, e.g. Blum *et al.* (2001) for the IALS and Goldstein (2004) for the PISA survey. But this remains an underresearched area. Our second contribution is to show the extent to which the cross-national pattern of results from one survey changes with the use of two variants of a standard item response model.

In both contributions we focus on two substantive issues. The first is the cross-country pattern of central tendency and of dispersion. How well children and young people in any country are doing on average is important to know in a globalized world. We also need to measure the educational inequalities within each country that help to generate differences in incomes and other aspects of living standards in later life. In both cases the performance of other countries is one natural yardstick.

The second issue is the relationship of central tendency to dispersion, which is also a topic of natural interest. Do the various surveys and scaling methods provide a clear picture of the association of these two basic features of score distributions? For example, do they suggest a trade-off between higher mean achievement and lower dispersion?

Section 2 introduces the four surveys that we consider, focusing on why results might differ between them. Section 3 compares results from these surveys. Section 4 investigates the robustness of results to choice of item response model. We concentrate on the 1995 TIMSS where results based on two different models are available from the survey organizers but we also discuss implications for comparisons across surveys. Section 5 concludes.

2. The international achievement surveys

Table 1 lists the data that we use. The PIRLS, TIMSS and PISA surveys collect data on school-age children. Schools are sampled (with probability proportional to size) and then a whole class (TIMSS and PIRLS) or a sample of all pupils (PISA) is randomly selected within each school. Sample size averages about 4000–6000 children per country, depending on the survey. By contrast, the IALS is a household survey of people of working age; we restrict attention to young people aged 16–24 years, of whom there are on average about 700 per country. For the TIMSS survey, we use data from both 1995 and 1999 rounds, taking the earlier year if a country did not participate in the later round. (See Appendix A for details.)

Country coverage varies from survey to survey. Section 3 concentrates on 18 countries that are present in the TIMSS, PISA and IALS and on 21 in the TIMSS, PISA and PIRLS surveys. The first group is composed of Organisation for Economic Co-operation and Development (OECD) members, i.e. countries at broadly similar levels of national income. Hence cross-

Table 1. Cross-national survey data used in the paper†

Survey	*Round*	*Age group (years)*	*Subjects covered*	*Average sample size per country*
TIMSS	1995 1999	13–14 (grade 8)	Mathematics and science	3800
PISA	2000	15	Reading, mathematics and science	5700
IALS	1994–1998	16–24	Document, prose and quantitative literacy	700
PIRLS	2001	9–10 (grade 4)	Reading	4300

†The first round of the PISA survey in 2000 was repeated in several further countries in 2002 in 'PISA+'. Several new entrants to the Organisation for Economic Co-operation and Development covered by the PISA+ survey are included in our analysis.

country differences are not driven by factors that are associated with large differences in development level. The second group contains 14 OECD members, two other rich countries (Hong Kong and Israel) and five Central and Eastern European countries at lower levels of development (Russia, Latvia, Bulgaria, Macedonia and Romania). Section 4 uses all 39 countries in the 1995 TIMSS for which microdata are available, of which only 24 are from the OECD. The distinction between rich and poor countries turns out to be important for the sensitivity of results to choice of item response model.

There are three sets of reasons why results may differ from survey to survey. First, the surveys aim to measure different things. Second, they all suffer from sampling and non-sampling errors in ways that may vary across surveys. Third, they may use different item response models.

2.1. Measurement aims

2.1.1. Subject

The surveys collect information on performance in various subjects. A country may perform well in one subject owing to a traditional emphasis in the area concerned, but less well in another. The TIMSS and PISA surveys both cover mathematics and science. The PISA survey in addition covers reading, which is the (sole) focus of the PIRLS survey. The IALS measures 'quantitative', 'prose' and 'document' literacy; the first uses a mathematical skill (essentially arithmetic) whereas the second requires reading skills. For convenience we refer to all four surveys as measuring 'achievement' in the subjects covered and to the assessment of each subject in a survey as a 'test'. Hence we have information on achievement from eight tests for the 18 countries in the TIMSS, PISA and IALS surveys and from six tests for the 21 countries in the TIMSS, PISA and PIRLS surveys. In contrast with some researchers (e.g. Brown (1999)) we do not disaggregate into different aspects of each subject within each survey.

2.1.2. Type of achievement

There are differences across surveys in type of achievement assessed, which again may cause the cross-country picture to vary. The IALS focuses on literacy skills that are needed for everyday tasks, e.g. working out a tip, calculating interest on a loan and extracting information from a timetable. The PISA survey also emphasizes knowledge to address real life settings with similarities to the IALS conceptual approach (Organisation for Economic Co-operation and Development (2001), page 18). By contrast, the TIMSS survey measures mastery of interna-

tionally agreed curricula and there is variation in how these match individual countries' actual curricula in mathematics or science. It is less clear how the PISA and PIRLS surveys differ in approach to reading. PIRLS organizers argue that the approaches are similar, both being based on 'an expanded notion of literacy' (Campbell *et al.* (2001), page 85).

2.1.3. Age group

The PIRLS survey covers young children. PISA and TIMSS children are in their early or mid-teens. Our IALS results relate to young people who were aged 16–24 years. Countries may do well at one age and not at another. One difference across surveys in age coverage is more subtle. The PISA survey targets children of a given age, whereas the TIMSS and PIRLS surveys cover children in a school 'grade'. Some countries promote all children at the end of the year to the next grade irrespective of achievement, whereas others insist on a certain competence before allowing passage upwards. Where the latter practice exists, average achievement relative to other countries can be expected to be higher in the TIMSS than in the PISA survey. But the same countries might show higher disparities in achievement in the PISA survey.

2.1.4. Calendar year

The surveys differ in the year for which they aim to measure achievement. Data collection in the various rounds of the surveys that we use span 1994–2001. Some change in the distribution of achievement is possible over such a time span and it could be different across countries.

2.2. Sampling and non-sampling errors

2.2.1. Sampling variation

Even if the surveys were to be identical in every aspect of design (target population, sampling scheme, test subjects, survey instrument etc.), sampling error would imply that their patterns of results would not correlate perfectly. Their results would be based on different samples of individuals. In practice, sampling error can be expected to be more of an issue for measures of dispersion than for central tendency, since the latter is easier to measure well.

2.2.2. Response

The surveys all suffer from non-response. Among the 21 countries in the TIMSS, PISA and PIRLS surveys, overall response (taking into account both school and student levels) averaged 83% for the PISA, 89% for the TIMSS and 90% for the PIRLS survey. Response to the IALS (in all countries) from working-age adults averaged 63%. The correlation in the country response rates between surveys is positive but not that high: 0.51 for PISA–TIMSS, 0.38 for PISA–PIRLS and 0.42 for TIMSS–PIRLS. Non-response bias affecting estimates of central tendency or dispersion for any country is unlikely to be the same across surveys.

2.2.3. Language and cultural bias

There are well-known difficulties in producing test instruments in international surveys that are culturally and linguistically neutral (Harkness *et al.*, 2002). Organizers of the achievement surveys put considerable effort into this area but inevitably there are concerns that full comparability is not obtained. For example, Blum *et al.* (2001) made a critical comparison of the French language IALS questionnaire that was used in France with the version that was used in Switzerland. (France originally participated in the IALS but later withdrew.) There is no reason

to believe that this source of measurement error is the same for a country in each survey given the differences in the subjects that are covered and the type of achievement that is assessed.

2.2.4. Detail and form of testing

Surveys cover the same subject area in differing degrees of detail. The TIMSS and PISA surveys both assess mathematics and science. But the 1999 TIMSS mathematics and science assessments had about 150 items compared with about 30 for these subjects in the 2000 PISA survey which in that year concentrated on reading, with the assessment of mathematics and science taking second place. There are differences in the form of testing also. About two-thirds of the 1999 TIMSS questions were multiple-choice questions, significantly more than in the PISA survey. Only about a third of the PIRLS assessment (in terms of possible scores) is based on this form of test. The IALS has no multiple-choice element. Arguably children in some countries do better at multiple-choice questions than children in others because of variation in countries' traditions of this form of testing in schools (e.g. O'Leary (2002)).

2.3. Item response models

A respondent's answers are summarized into a single score for the subject concerned—mathematics, science, reading, different types of literacy, etc. We defer discussion of this procedure to Section 4 but one aspect needs to be dealt with here before we compare results across surveys in Section 3. For each test, scores are scaled to produce values that are chosen by the organizers for the mean and standard deviation among all the people in participating countries—500 and 100 respectively in subjects in the TIMSS, PISA and PIRLS surveys, and about 275 and 50 in the IALS. None of the scores are directly comparable across surveys because the overall mean and standard deviation in each case are based on a different group of countries. The TIMSS and PIRLS surveys both include a wider range of countries in terms of development level than does the PISA survey, which covered OECD members only in 2000. So, for example, that Italy had a mean reading score of 541 in the PIRLS but only 487 in the PISA survey in part reflects the fact that the PIRLS survey included such countries as Belize, Columbia and Morocco whereas the PISA scale is based solely on OECD countries.

We use two methods to overcome this problem. First, within each of the two groups of countries that are present in three surveys, we compare country *rankings* across the tests concerned. Rankings have the advantage of being easily understood and compared. They have the disadvantage of ignoring all information on the extent of differences between countries. And, inevitably, they suggest that national performance is like a beauty parade where coming first is all important. Our use of rankings is not intended to propagate that view—we rank to compare more easily across tests. Second, we convert the measures of central tendency and dispersion for each country into *z-scores*, i.e. for the pool of 18 countries in the PISA, TIMSS and IALS and the 21 in the PISA, TIMSS and PIRLS surveys, we adjust the measure concerned (e.g. each country's median) by subtracting the mean value for the pool in question and by dividing by the standard deviation of the values for that pool. (Appendix B gives examples.) Country rankings and correlations between the country values are unchanged by this transformation.

In all three areas—measurement aims, sampling and non-sampling errors, and item response models—there are reasons why the cross-country pattern of results may vary from survey to survey. This means that we cannot rely on a single test for an adequate picture of a country's educational achievement. Our aim is to establish the extent of the variation in results from test to test and, in the case of item response modelling, to pinpoint the contribution that is made by the choice of model.

3. Comparing results across surveys

Do different surveys and subjects give a similar picture of country differences in central tendency and dispersion? We measure central tendency by the median and dispersion by the difference between the 95th and fifth percentiles, P95–P5 (the results are not sensitive to these choices).

Fig. 1 gives a graphical summary that includes all eight tests in the PISA, TIMSS and IALS for the 18 countries that were covered by these surveys. It plots each country's *average rank* for the median against that for P95–P5. Each country's value of the median or P95–P5 is ranked for each test and the average values of its ranks are calculated, weighting the surveys equally (rather than the tests). (Appendix B gives details.) These average ranks have considerable merit as summary statistics. If the different tests produced wildly differing rankings then the averaging would produce figures with little variation. A low rank in one test would be likely to be balanced by a high rank in another, leaving all 18 countries clustered around an average rank of 9.5. The more the average ranks vary the more the separate rankings for each test must be in agreement.

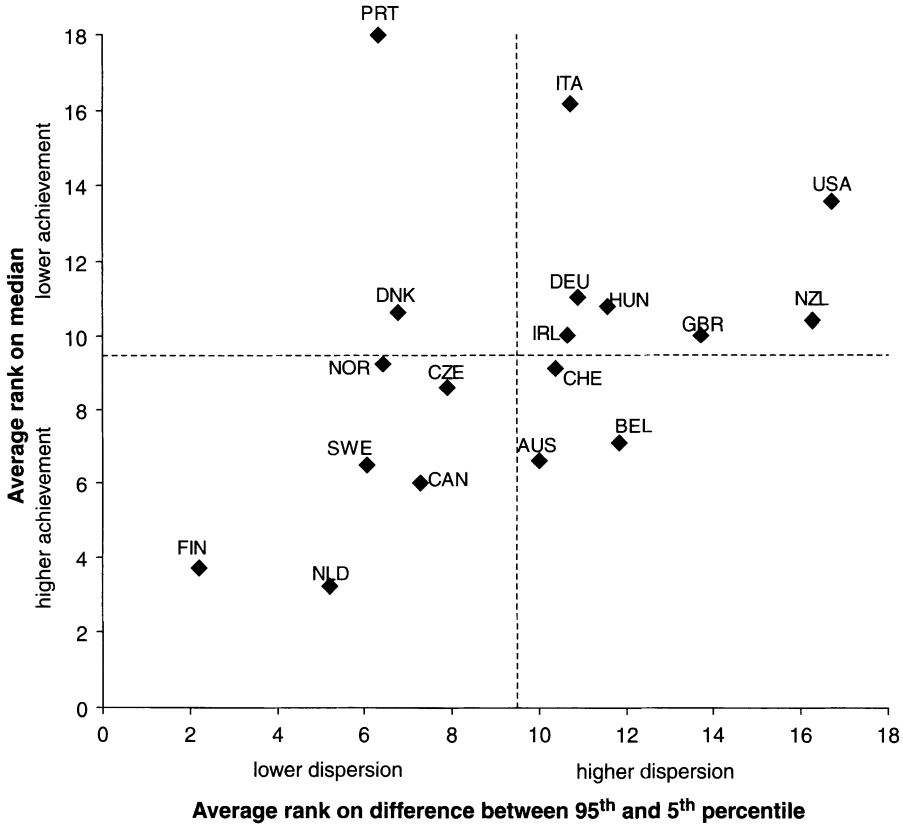

Fig. 1. Average ranks on central tendency (median) and dispersion (P95–P5) for 18 countries in eight tests (PISA, TIMSS and IALS): the higher the median and the lower the dispersion (P95–P5) the smaller in number the rank; grid lines show the average for all countries (9.5) (PRT, Portugal; ITA, Italy; DNK, Denmark; DEU, Germany; HUN, Hungary; IRL, Republic of Ireland; GBR, UK; NZL, New Zealand; NOR, Norway; CZE, Czech Republic; CHE, Switzerland; AUS, Australia; BEL, Belgium; CAN, Canada; SWE, Sweden; FIN, Finland; NLD, the Netherlands)

Having a low or high average rank can only result from ranking consistently well or consistently badly in each survey. ('Well' means a higher value of the median than other countries or a smaller value of P95–P5.)

Three features of the results stand out. First, the average ranks display considerable variation. Our first substantive question that was outlined in Section 1 was whether the various surveys give a similar cross-country picture of central tendency and dispersion. The variation in average ranks is encouraging evidence for a positive answer. However, it is also true that there is bunching in the middle of the distribution on each measure, arising either from countries consistently ranking mid-table or from an evening-out of good performance on one test and bad performance on another.

Second, a higher average rank on the median tends to be associated with a higher rank on P95–P5. Countries with higher average achievement have, in general, smaller within-country differences. This starts to answer our second substantive question, which is on the relationship between central tendency and dispersion.

Third, several countries are in obvious extreme positions or are outliers. Finland has an average rank of only 3.7 on the median and 2.2 on P95–P5. At the opposite end of the spectrum the USA averages 13.6 and 16.7 respectively on the two measures. Italy and Portugal stand out as exceptions to the general pattern of association between central tendency and dispersion. Despite mid-table and high table positions respectively on dispersion (in average rank terms) they rank very lowly on the median. Indeed, Portugal has the lowest median score in all eight tests and hence an average rank of 18.

Tables 2 and 3 shed more light on how the average ranks come about for the median and P95–P5 respectively, showing the country z-scores for each test. The shading in the 4th–11th columns indicates the third of the distribution for that test in which a country falls: dark shading for the lowest third, light shading for the middle third and white for the top third. The countries are ordered on the basis of the average ranks that are used in Fig. 1. The values of these averages are given in the second column and the average z-scores (again weighting surveys equally) are given in the third column.

Both Finland and the Netherlands have medians that on average are more than 1 standard deviation above the group mean. Portugal, at the other extreme, averages 2 standard deviations below the mean. In the middle of the distribution, the UK's average rank of 10.1 reflects a considerable mix of results for individual tests. Whereas all the UK's PISA z-scores are positive, all those for the IALS are negative, showing a clear difference between the two surveys. This mix of results is found for quite a few other countries as well: a half of all countries have three different shades in their row of entries.

Table 3, relating to dispersion, also has half of the countries with this pattern of results. Germany is an interesting case of disagreement between the results of the PISA and the other two surveys. The high dispersion in PISA scores in Germany has been much commented on (e.g. Baumert *et al.* (2001)) whereas the IALS shows dispersion for 16–24-year-old Germans to be among the smallest for the 18 countries.

Fig. 2 switches to the 21 countries that were covered by the PISA, TIMSS and PIRLS surveys, again showing average ranks for the median and for P95–P5. This comparison replaces the 16–24-year-olds in the IALS with the youngest age group covered by any of our four sources, the PIRLS 10-year-olds. The PIRLS survey covers just one subject, reading, and we again weight surveys equally when combining results across tests (so the PIRLS ranks contribute a third of the average ranks). Of course, the average ranks for any country must be interpreted in relation to the pool of countries, which has now changed from that in Fig. 1.

Table 2. Average ranks and *z*-scores for the median for 18 countries in eight tests (PISA, TIMSS and IALS)†

Country	*Average rank*	*Average z-score*	*Results for the following surveys:*							
			PISA			*TIMSS*		*IALS*		
			Reading	*Mathematics*	*Science*	*Mathematics*	*Science*	*Document*	*Quantitative*	*Prose*
Netherlands	3.2	1.11	1.11	1.97	1.12	1.26	1.26	0.77	0.50	0.72
Finland	3.7	1.05	1.69	0.81	1.36	0.38	0.63	1.31	0.71	2.02
Canada	6.0	0.63	1.08	0.71	1.00	0.76	0.54	0.55	0.05	0.35
Sweden	6.5	0.63	0.32	−0.03	0.33	0.13	0.17	1.34	1.43	1.84
Australia	6.6	0.51	0.83	0.73	0.92	0.63	0.97	−0.11	−0.19	0.03
Belgium	7.1	0.41	0.32	0.62	−0.11	1.29	−0.54	0.55	0.77	0.41
Czech Republic	8.6	0.16	−0.84	−0.56	0.12	0.15	0.75	0.36	1.18	−0.16
Switzerland	9.1	0.19	−0.62	0.67	−0.62	1.12	−0.26	0.52	0.55	−0.10
Norway	9.2	0.12	−0.06	−0.40	−0.22	−0.50	−0.22	1.14	0.76	0.99
UK	10.1	0.03	0.51	0.60	1.12	−0.73	0.63	−0.37	−0.93	−0.46
Ireland	10.1	−0.02	0.78	−0.23	0.24	0.45	0.04	−0.85	−0.59	−0.29
New Zealand	10.4	−0.05	0.99	0.95	1.06	−0.85	−0.30	−0.60	−0.85	−0.23
Denmark	10.7	−0.36	−0.53	0.12	−1.12	−0.44	−1.95	0.90	0.97	0.05
Hungary	10.8	−0.22	−1.39	−0.93	−0.50	0.90	1.51	−1.16	−0.14	−1.45
Germany	11.1	−0.22	−1.00	−0.61	−0.84	−0.27	0.22	0.08	0.58	−0.11
USA	13.6	−0.67	−0.25	−0.62	−0.33	−0.39	−0.10	−1.20	−1.77	−1.12
Italy	16.2	−1.19	−1.09	−1.83	−1.30	−1.29	−1.16	−1.13	−1.07	−0.61
Portugal	18.0	−2.13	−1.85	−1.96	−2.23	−2.60	−2.19	−2.12	−1.95	−1.89

†Surveys equally weighted.

Table 3. Average ranks and *z*-scores for P95–P5 for 18 countries in eight tests (PISA, TIMSS and IALS)†

Country	*Average rank*	*Average z-score*	*Results for the following surveys:*							
			PISA			*TIMSS*		*IALS*		
			Reading	*Mathematics*	*Science*	*Mathematics*	*Science*	*Document*	*Quantitative*	*Prose*
Finland	2.2	−1.35	−1.45	−1.76	−1.68	−1.70	−0.87	−0.99	−1.47	−0.93
Netherlands	5.2	−0.75	−1.38	−0.94	−0.02	−0.51	−1.01	−0.92	−0.47	−0.79
Sweden	6.1	−0.60	−0.86	0.15	−0.68	−1.08	−1.05	−0.42	−0.38	−0.01
Portugal	6.3	−0.72	−0.18	−0.27	−1.41	−1.65	−1.29	−0.23	−0.47	0.47
Norway	6.4	−0.47	0.72	−0.12	−0.22	−0.40	−1.37	−0.47	−0.52	−0.95
Denmark	6.8	−0.52	−0.19	−0.96	0.93	−0.19	0.04	−1.29	−1.12	−1.78
Canada	7.3	−0.48	−0.58	−1.17	−1.27	−0.62	−0.86	0.82	−0.09	0.17
Czech Republic	7.9	−0.28	−0.23	0.62	−0.42	0.23	−0.53	−0.22	−0.81	−1.04
Australia	10.0	0.03	0.34	−0.30	−0.44	0.28	0.41	−0.39	−0.05	0.04
Switzerland	10.4	−0.04	0.50	0.96	0.42	−1.15	−0.52	0.14	−0.08	0.20
Ireland	10.7	0.13	−0.65	−1.38	−0.83	0.75	0.88	0.35	0.83	0.40
Italy	10.7	0.14	−1.19	−0.28	0.13	1.22	0.56	−0.25	0.06	0.15
Germany	10.9	0.33	1.88	1.38	0.95	−0.06	1.13	−1.10	−1.14	−0.63
Hungary	11.6	0.36	−0.79	0.68	0.78	1.10	0.02	0.61	0.60	−0.36
Belgium	11.8	0.55	1.21	1.88	2.41	0.21	0.08	−0.85	−0.09	−0.05
UK	13.7	0.80	0.28	−0.15	0.27	0.70	1.13	1.32	1.63	1.08
New Zealand	16.3	1.28	1.43	0.81	0.48	1.53	1.31	1.52	1.24	1.77
USA	16.7	1.61	1.15	0.83	0.58	1.35	1.94	2.38	2.34	2.25

†Surveys equally weighted.

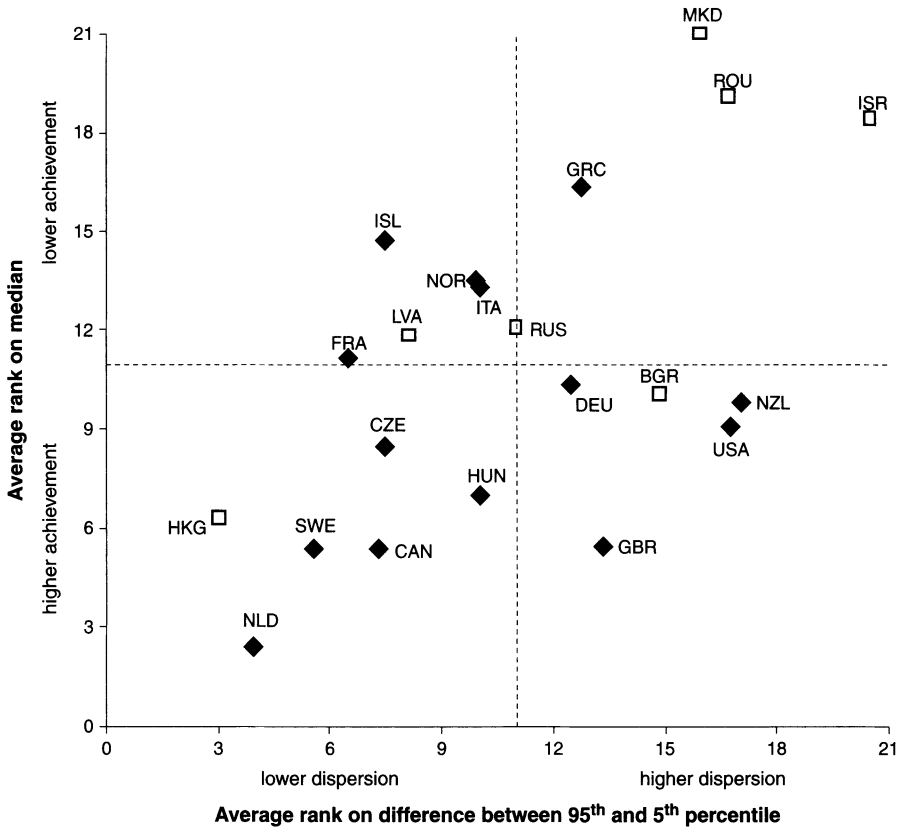

Fig. 2. Average ranks on the median and P95–P5 for 21 countries in six tests (PISA, PIRLS and TIMSS): the higher the median and the lower the dispersion (P95–P5) the smaller in number the rank; grid lines show the average for all countries (♦, OECD countries; □, other countries) (MKD, Macedonia; ROU, Romania; ISR, Israel; GRC, Greece; ISL, Iceland; NOR, Norway; ITA, Italy; LVA, Latvia; RUS, Russia; FRA, France; DEU, Germany; BGR, Bulgaria; NZL, New Zealand; CZE, Czech Republic; HUN, Hungary; HKG, Hong Kong; SWE, Sweden; CAN, Canada; GBR, UK; NLD, the Netherlands)

The new countries of Macedonia, Romania and Israel stand out as having low average achievement and high dispersion. Hong Kong in contrast has the smallest within-country differences of any country. These are clear results, both for the countries that are concerned and in terms of re-enforcing the pattern of association between central tendency and dispersion in Fig. 1: on average within-country differences are lowest where average scores are highest.

The move to a group of countries that includes some notable weak performers from outside the OECD means that the UK's relative position improves for both central tendency and dispersion. As far as the median is concerned, the same effect is produced by the replacement of the results of the IALS, in which the UK performed badly, with the results of the PIRLS survey where the UK did well. However, on dispersion the UK once again stands out in the PIRLS survey as a country with high within-country differences. The situation is similar for the USA and New Zealand: their relative positions improve on both the median and P95–P5 owing to the change in the country pool but the substitution of the PIRLS for the IALS replaces one survey in which the dispersion of their scores is high for another where the same is true. The partial changes in the pools of tests and countries between Figs 1 and 2 does not change the conclusion that these three countries have large within-country differences by international standards.

Table 4. Correlation matrix of the medians for 18 countries covered by PISA, TIMSS and IALS

Statistic	*Survey*		*Results for the following surveys:*							
			PISA			*TIMSS*		*IALS*		
			Reading	*Mathematics*	*Science*	*Mathematics*	*Science*	*Prose*	*Document*	*Quantitative*
Median	PISA	Reading	1							
		Mathematics	0.82	1						
		Science	0.90	0.80	1					
	TIMSS	Mathematics	0.46	0.65	0.52	1				
		Science	0.44	0.47	0.72	0.66	1			
	IALS	Prose	0.67	0.57	0.57	0.43	0.27	1		
		Document	0.50	0.61	0.46	0.54	0.25	0.91	1	
		Quantitative	0.21	0.40	0.24	0.59	0.28	0.74	0.89	1
P95–P5	PISA	Reading	1							
		Mathematics	0.73	1						
		Science	0.57	0.73	1					
	TIMSS	Mathematics	0.31	0.33	0.50	1				
		Science	0.51	0.33	0.47	0.80	1			
	IALS	Prose	0.37	0.28	0.05	0.47	0.60	1		
		Document	0.25	0.17	0.00	0.56	0.55	0.87	1	
		Quantitative	0.28	0.23	0.23	0.70	0.67	0.88	0.91	1

One disadvantage of the average ranks and z-scores is the equal weight that is given to an agreement between tests within the same survey and to an agreement between tests in different surveys. (Given our equal weighting of surveys rather than tests, this is only strictly true when the number of tests per survey is equal, as in the PISA and IALS.) We may well want to take more notice of the latter: agreement across surveys. This motivates analysis of the correlations between the z-scores for each pair of tests, which are given in Tables 4 and 5 for both the 18-country and the 21-country groups. Are the correlations within survey for different subjects higher than those between surveys for similar subjects? The answer is 'yes' in Table 4: the within-survey correlations are higher than almost every correlation between tests in different surveys, and this is true for both central tendency and dispersion. The same pattern is also found in Table 5 where the inclusion of countries at lower levels of development pushes up the within-survey correlations of country scores in the PISA and TIMSS surveys.

But it is also true that, among the correlations between tests from different surveys, the values for subjects that are similar are typically *higher* than those for other subjects. This encourages confidence in the general message to be obtained about a subject from each survey.

The correlations for P95–P5 are in general lower than for the median: there is more agreement between tests on the country pattern of central tendency than for dispersion. This does not seem surprising, the latter being harder to measure well. And, as we shall see in Section 4, the measurement of dispersion appears to be much more sensitive to the choice of item response model, which may differ from survey to survey.

We undertook two sensitivity analyses for the correlations between tests (see Brown *et al.* (2005) for details). The first concerns the age of respondents. Correlations between test results in the TIMSS and PISA surveys might be expected to be higher (*ceteris paribus*) than those between either survey and the PIRLS or IALS on account of the similarity in the ages of children

Table 5. Correlation matrix of the median and P95–P5 for 21 countries covered by PISA, TIMSS and PIRLS

Statistic	*Survey*		*Results for the following surveys:*					
			PISA			*TIMSS*		*PIRLS, reading*
			Reading	*Mathematics*	*Science*	*Mathematics*	*Science*	
Median	PISA	Reading	1					
		Mathematics	0.94	1				
		Science	0.96	0.96	1			
	TIMSS	Mathematics	0.58	0.72	0.67	1		
		Science	0.59	0.66	0.70	0.73	1	
	PIRLS	Reading	0.58	0.51	0.57	0.50	0.68	1
P95–P5	PISA	Reading	1					
		Mathematics	0.56	1				
		Science	0.57	0.63	1			
	TIMSS	Mathematics	0.42	0.71	0.35	1		
		Science	0.58	0.68	0.46	0.89	1	
	PIRLS	Reading	0.48	0.39	0.13	0.65	0.68	1

who were covered. However, the PISA study surveys children of a given age whereas the TIMSS survey targets a school grade. Section 2 noted possible consequences for a comparison of results from the two sources. To try to adjust for the difference in approach, we recalculate PISA–TIMSS correlations using subsamples of children of the same age from the TIMSS and of the same grade from the PISA survey. The effect is to raise somewhat the correlations for values of P95–P5 for both the 18- and the 21-country pools that are covered by Tables 4 and 5. However, there are mixed effects for the median correlations.

The second issue is the effect of sampling error. In practice sampling error is more of an issue for P95–P5 than for the median. We use published information on standard errors in the TIMSS, PISA and PIRLS surveys to estimate the effect of sampling error on the Table 5 correlations. We estimate that the correlations between the median values in different surveys would typically increase only very slightly if sampling error were eliminated completely. However, the correlations for P95–P5 would rise by an average of 0.07. This is sufficient to close much of the difference between the average (off-diagonal) levels of correlation for central tendency and dispersion.

Three conclusions come from the comparisons in this section. First, there is considerable agreement on both central tendency and dispersion between the various tests that are contained in the four surveys, as summarized by average ranks and z-scores. This agreement is sufficient to establish a general pattern of association between the two aspects of the distributions, with higher average scores and smaller within-country differences tending to go together. Second, care is nevertheless needed in judging the record of individual countries, with the different subjects and surveys quite frequently giving rather different results. Third, agreement between tests in different surveys tends to be less than agreement between tests within the same survey. Among other things, this underlines the importance of considering factors that may be peculiar to each survey. These include the item response modelling, which is the subject of the next section.

4. Comparing item response models

Item response models are used by the survey organizers to produce summary scores for each individual. These scores are *derived* data and the question arises whether the choice of method of derivation has an influence on the results. Too little is known about this. Typically nothing is said on the subject in the survey reports. Many users access only those published sources. Even where secondary analysis is made of the microdata, the procedures that are involved in fitting the models are sufficiently complex that it is impractical for most researchers to try alternatives.

We see how estimates of central tendency, dispersion and the association between the two change for one survey, the 1995 TIMSS, when two different item response models are applied to the data. This isolates the effect of model choice. We then comment on the implications for differences in results across surveys given the type of item response model that each survey organizer uses.

Models that are employed by survey organizers are invariably 'unidimensional', which is appropriate when high ability individuals have a greater probability than low ability individuals of answering each and every question correctly. Goldstein (2000, 2004) criticized this assumption, experimenting with less restrictive 'multidimensional' models. We confine attention to unidimensional models to explore robustness *within* this class of model. Like Goldstein, we are concerned with the sensitivity of results to modelling choices.

The unidimensional models that are applied by survey organizers are typically 'one-parameter' or 'three-parameter' logit models. The purpose in both cases is to estimate a person's 'proficiency' in a subject (mathematics, science, etc.) from answers to a number of questions. The one-parameter model allows for differences in the difficulty of each question. The three-parameter model allows also for the probability that the answer is guessed and for a question's ability to discriminate between students of high and low proficiency. These models give the probability of a correct answer to question i by student j as, for the one-parameter model,

$$p_{ij}(\text{correct answer}) = \frac{1}{1+\exp\{-(\theta_j - \alpha_i)\}}$$

and, for the three-parameter model,

$$p_{ij}(\text{correct answer}) = \gamma_i + \frac{1-\gamma_i}{1+\exp\{-\beta_i(\theta_j - \alpha_i)\}}$$

where θ_j is a student's proficiency, α_i is a question's difficulty, γ_i is the probability that the answer to a question is guessed and β_i measures the power of a question to discriminate between individuals of high and low ability. The estimation of a logit model, in which the θ_j are treated as unobserved fixed effects to estimate the other parameters, is only the first step in the derivation of the scores. The logit functional form is just one of several alternatives for modelling the probability of a correct answer; Goldstein (1980) compared results from a logit model and a complementary log–log-model, noting their differences in treatment of high and low ability. We do not pursue this aspect of robustness here.

Results for the 1995 TIMSS have been produced by the survey's organizers with both types of model. A one-parameter model was used for the survey reports (Beaton *et al.*, 1996a, b). The three-parameter model that was used for the 1999 TIMSS was also applied to the 1995 data to allow results to be compared over time. (Where 1995 data are used in Section 3, the results are from the 'three-parameter' model.) No systematic analysis appears to have been published of differences in results from the two sets of scores. However, the 1995 microdata that were

derived from the three-parameter model are available for each country on the TIMSS Web site `www.timss.org` alongside the data that are based on the one-parameter model (including for those countries which are not in the 1999 survey). These two sets of microdata are the basis for our analysis and are available for 39 of the 40 countries that were covered by the 1995 TIMSS (the exception is Italy). We refer to the two sets of scores as one-parameter scores and three-parameter scores, although there is another difference between them: at an intermediate stage in the process of deriving the latter, θ was modelled as a function of characteristics of the student and his or her school.

Fig. 3 shows the distributions of the two sets of mathematics scores that were derived from the same raw data for four countries, selected to illustrate the range of differences that occur. For the UK, the switch in item response model leads to a loss of positive skew but overall the distributions seem similar. The picture is not the same for the other three countries. For Singapore, there is a substantial reduction in dispersion. For Iran, there is a widening of the distribution, whereas for South Africa there is both a large reduction in the mean and a large increase in dispersion (and positive skew). We surmise that the changes in South Africa (and the smaller changes in other less developed countries) are due in particular to the three-parameter model's allowance for the probability of guessing. Controlling for guessing allows really poor ability to be better revealed, leading to a fall in the mean and a larger fall at the bottom of the distribution. A minority of children in South Africa have high achievement. Once the guessing probability is controlled for, the gap between these high performing children and those at the bottom of the distribution is revealed more clearly.

If distributions are changing in different ways from country to country we can expect that countries' standings relative to one another will change. We start with central tendency. Fig. 4 plots each country's median for the mathematics three-parameter scores against that for the one-parameter scores. To be clear: the raw data behind the two sets of scores—the answers that were given by respondents to the questions—are identical. What differs is the method that was used to summarize those data for each individual into a single number.

The conclusion seems straightforward. The medians are very highly correlated, both among OECD countries and among all countries covered by the 1995 survey. And this is true for both mathematics and science. The cross-country pattern of central tendency is robust to the change in item response model. However, for both subjects a few countries lie some way off the 45° line. South Africa (ZAF) is the most extreme case. There is a fall in the median for mathematics from the one- to the three-parameter scores of over 75 points (which is also shown clearly in Fig. 3). This is a big difference, changing the picture of just how far adrift the average South African child is from his or her counterpart in other countries.

We now turn to dispersion, which is measured as in Section 3 by the difference between 95th and fifth percentiles, P95 and P5. Fig. 5 shows what happens to each of these two quantiles, focusing on mathematics. (Similar results are found for science.) The correlations between one- and three-parameter scores are very high, as for the median. But, critically, the pattern of change for the two quantiles is not the same. For P5 the slope of a regression line would clearly be greater than 1 whereas for P95 it would be less than 1. For country values of P95–P5 to be highly correlated it is not sufficient that one- and three-parameter values for both quantiles display high correlation—the regression lines would also need to have the *same* slope.

The net result in terms of the change in P95–P5 is shown in Fig. 6 for both mathematics and science. For mathematics, the correlation between the two sets of values is essentially *zero* (0.03): in contrast with the median, the cross-country pattern of dispersion is therefore far from robust to the choice of item response model. (The correlations are very similar if the standard deviation is used in place of P95–P5.) The change in the position of South Africa is dramatic.

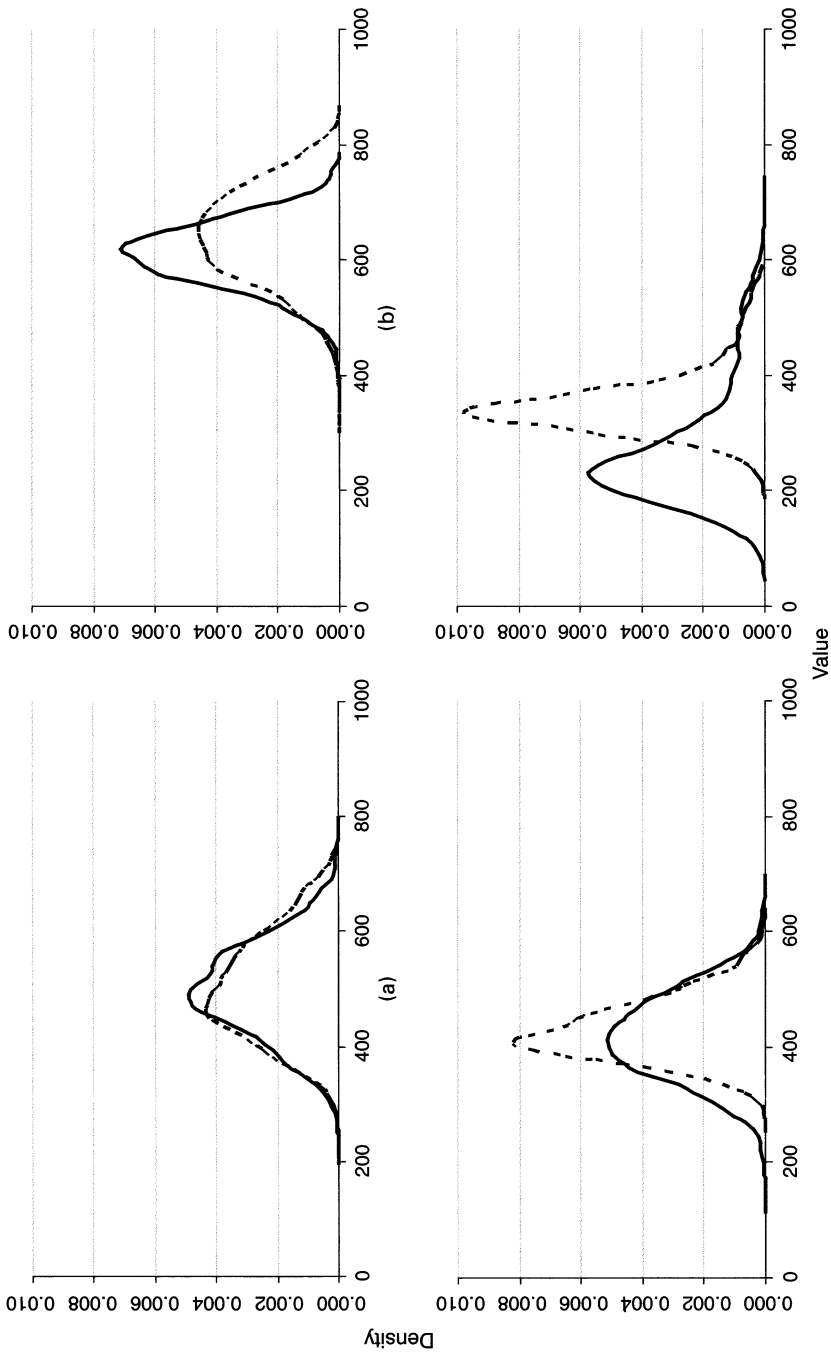

Fig. 3. Distribution of eighth-graders' achievement in mathematics in the TIMSS 1995 (the distributions that are depicted are of the averages of the five plausible values for each individual; - - - - - -, one-parameter scores; ——, three-parameter scores): (a) UK; (b) Singapore; (c) Iran; (d) South Africa

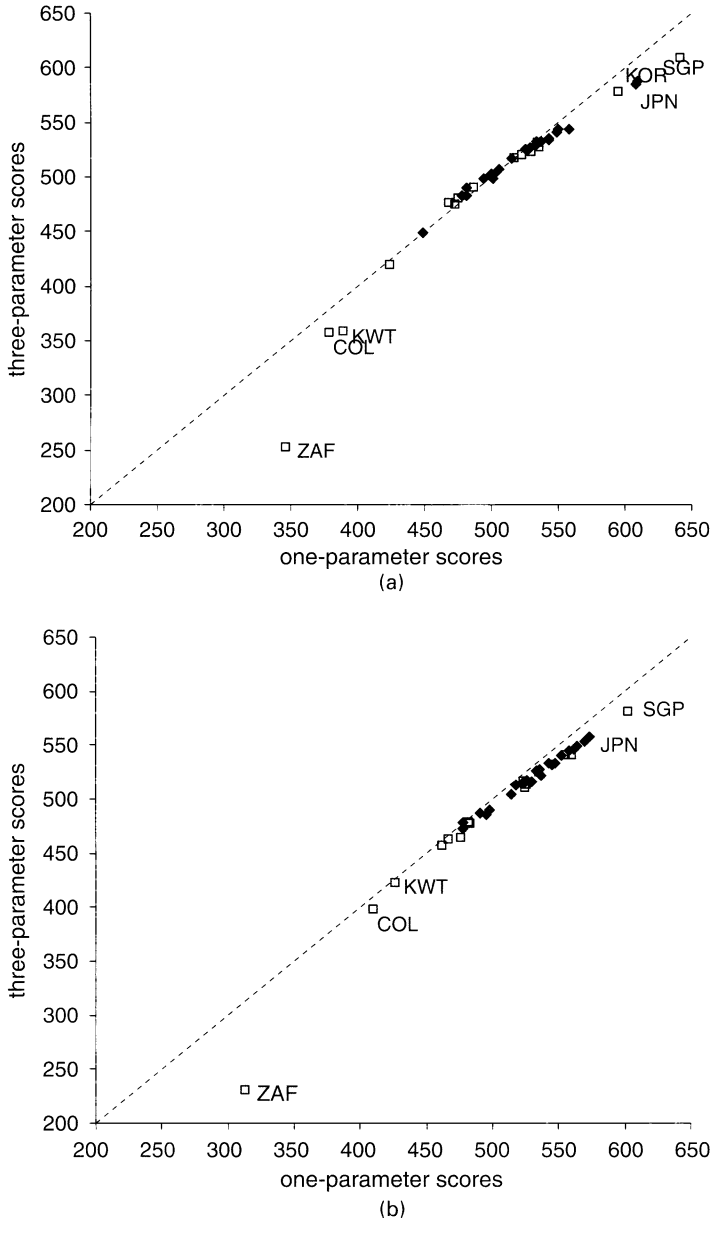

Fig. 4. One-parameter and three-parameter values of the median for the TIMSS 1995 (the correlations of the one- and three-parameter medians are 0.98 for mathematics (1.00 for OECD countries) and 0.97 for science (0.99 for OECD countries); ♦, OECD countries; □, other countries; KOR, Korea; SGP, Singapore; JPN, Japan; KWT, Kuwait; COL, Columbia; ZAF, South Africa): (a) mathematics; (b) science

The country with one of the smallest values for the one-parameter scores becomes the country with the greatest dispersion when judged by the three-parameter scores. The changes for Kuwait (KWT) and Columbia (COL) are almost as striking. Singapore (SGP), in contrast, changes from a middle ranking country for dispersion of one-parameter scores to the country with the smallest within-country differences in three-parameter scores.

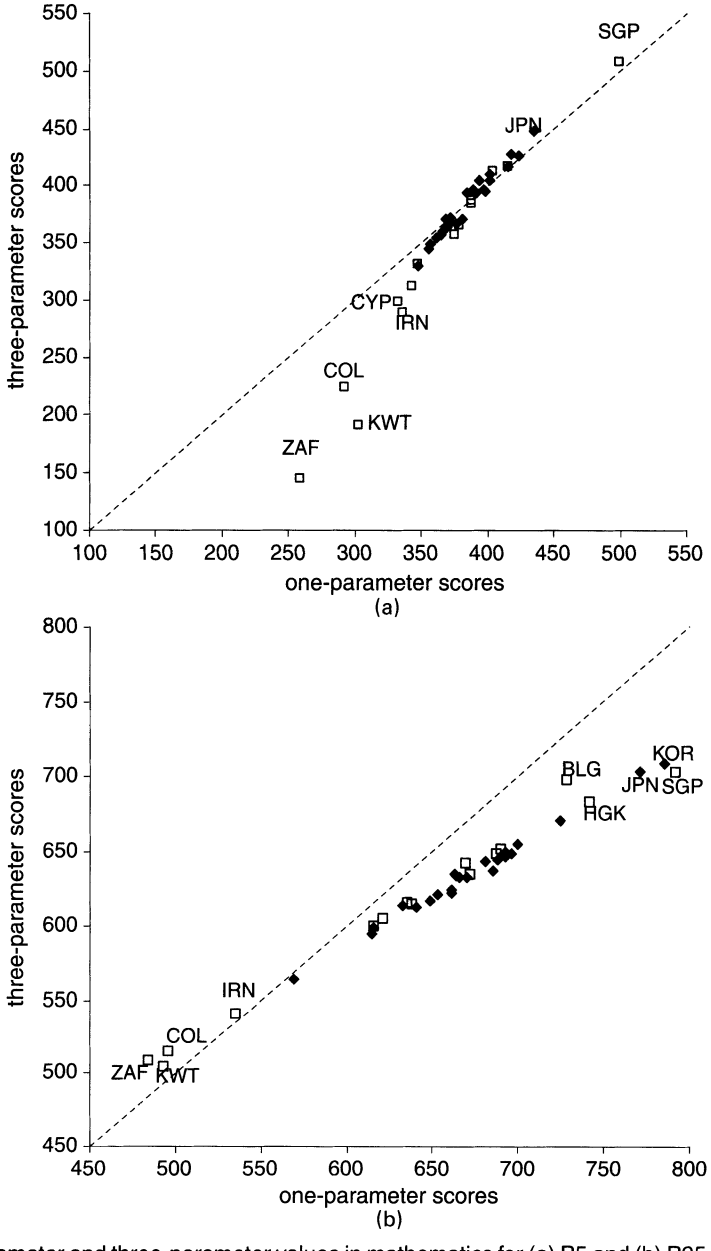

Fig. 5. One-parameter and three-parameter values in mathematics for (a) P5 and (b) P95, TIMSS 1995 (the correlations of the one- and three-parameter values are 0.97 for P5 (0.98 for OECD countries) and 0.99 for P95 (1.00 for OECD countries); ♦, OECD countries; □, other countries; SGP, Singapore; JPN, Japan; KOR, Korea; BLG, Bulgaria; HKG, Hong Kong; CYP, Cyprus; IRN, Iran; COL, Columbia; KWT, Kuwait; ZAF, South Africa)

The zero correlation is driven by the non-OECD countries. With these excluded the correlation rises to 0.70. The robustness of the ranking on dispersion is therefore much higher for these richer countries, which traditionally have been the core participants in the achievement surveys. However, even here some change is evident. For example, Greece (GRC) is at the OECD average for P95–P5 for the one-parameter scores but has the greatest dispersion in the OECD for the

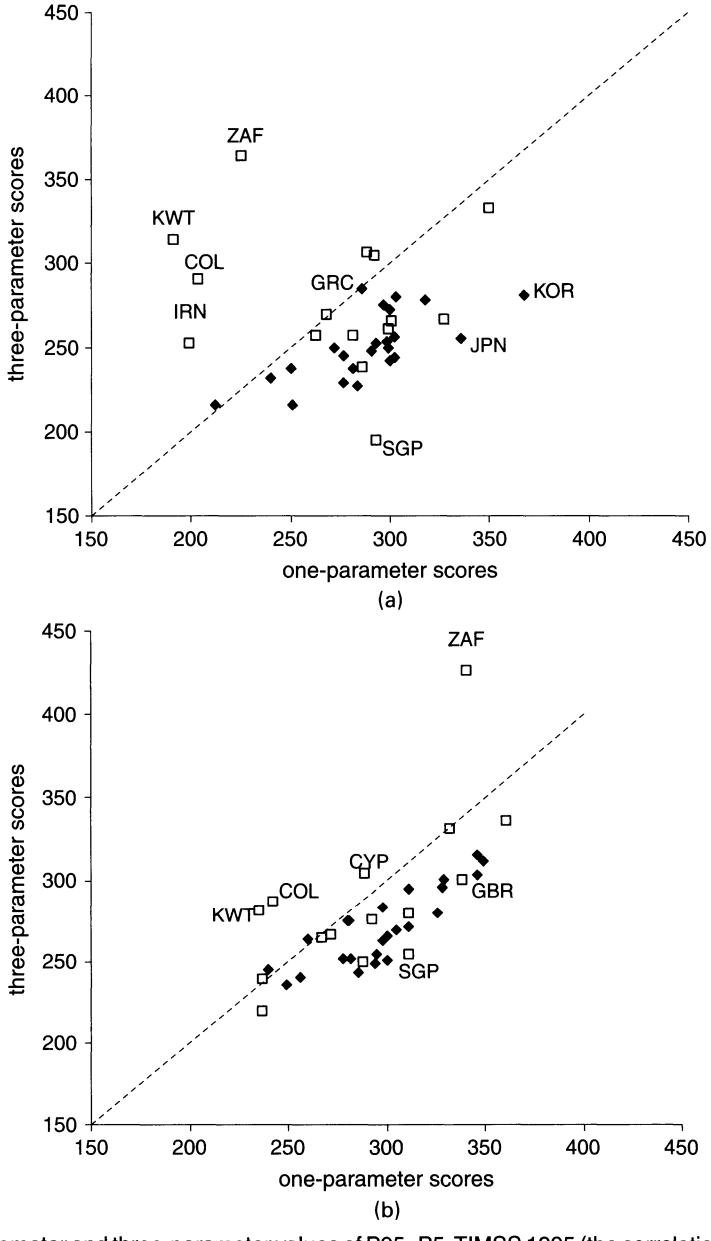

Fig. 6. One-parameter and three-parameter values of P95–P5, TIMSS 1995 (the correlations of the one- and three-parameter values of P95-P5 are 0.03 for mathematics (0.70 for OECD countries) and 0.67 for science (0.85 for OECD countries); ♦, OECD countries; □, other countries; ZAF, South Africa; KWT, Kuwait; COL, Columbia; CYP, Cyprus; GBR, UK; GRC, Greece; KOR, Korea; IRN, Iran; JPN, Japan; SGP, Singapore): (a) mathematics; (b) science

three-parameter scores. (Since Greece lies on the 45° line, this comes about from changes in the values for other countries.)

The change in item response model has much less effect for science. Nevertheless, there is still some notable reranking. For example, Kuwait and Columbia are again above the 45° line: dispersion of their three-parameter scores is now well above that in Singapore, rather than being

well below. With the one-parameter scores the UK (GBR) and Cyprus (CYP) are separated by 20 ranks whereas the dispersion in the two countries is almost identical for the three-parameter scores. South Africa becomes a big outlier, having been merely one of the countries with high dispersion of one-parameter scores.

Fig. 7 shows how the switch in item response model changes the view of whether dispersion rises or falls with central tendency, focusing on mathematics. With the one-parameter scores, countries with higher average achievement have higher dispersion in achievement ($r = 0.79$). With the three-parameter data the opposite conclusion would be drawn ($r = -0.58$). The latter was one of our conclusions from comparisons of surveys in Section 3 (where in the case of the TIMSS survey we used three-parameter data) although the focus there was mainly on OECD countries. If attention is restricted to those richer countries, then the change is not so sharp, the pattern changing from fairly strong to very weak positive correlation. The changes for science (which are not shown) are again less dramatic: weak positive correlation switching to weak negative correlation.

To summarize:

(a) the cross-country pattern of central tendency in the 1995 TIMSS is not sensitive to the choice of one- or three-parameter model;
(b) the pattern of dispersion for mathematics is quite sensitive with some sharp changes in country rankings that alter completely the picture of the outliers, but there is less sensitivity for the OECD countries and results for science also change much less;
(c) the direction of association of central tendency and dispersion for mathematics changes with the switch in item response model.

The greater sensitivity of results for less developed countries makes one wonder whether a single test instrument is suitable for such a wide range of countries in terms of average ability levels as are now included in the TIMSS survey.

What do these findings imply for comparisons of different surveys' results? The TIMSS results in Section 3 are all based on the three-parameter scores. Unless the item response model behind the results for the PISA, IALS and PIRLS data is the same as that for the TIMSS scores we were not comparing like with like.

The models that were used in the IALS and PIRLS analyses are similar to that for the three-parameter TIMSS scores: comparisons between any of these sources can rely on a high degree of comparability of model (see Brown *et al.* (2005) for details). However, the PISA analysis used a *one*-parameter model that was 'identical to that used in TIMSS 1995' (Adams (2003), page 386; see also Adams (2002)). As a consequence, the results in Section 3 for the PISA survey are not from the same type of item response model as those from the other surveys. Our findings in the present section show that this is very unlikely to make much difference to comparisons of central tendency, especially if the focus is restricted to the OECD countries. However, the greater sensitivity of measured dispersion to the choice of model suggests that comparisons of within-country differences in the PISA survey with those in the other surveys may potentially mislead.

To explore this we take mathematics score data for countries in both the 1995 TIMSS and the PISA surveys and compare correlations of central tendency (measured by the median) and dispersion (measured by P95–P5) between

(a) TIMSS three-parameter results and PISA results and
(b) TIMSS one-parameter results and PISA results.

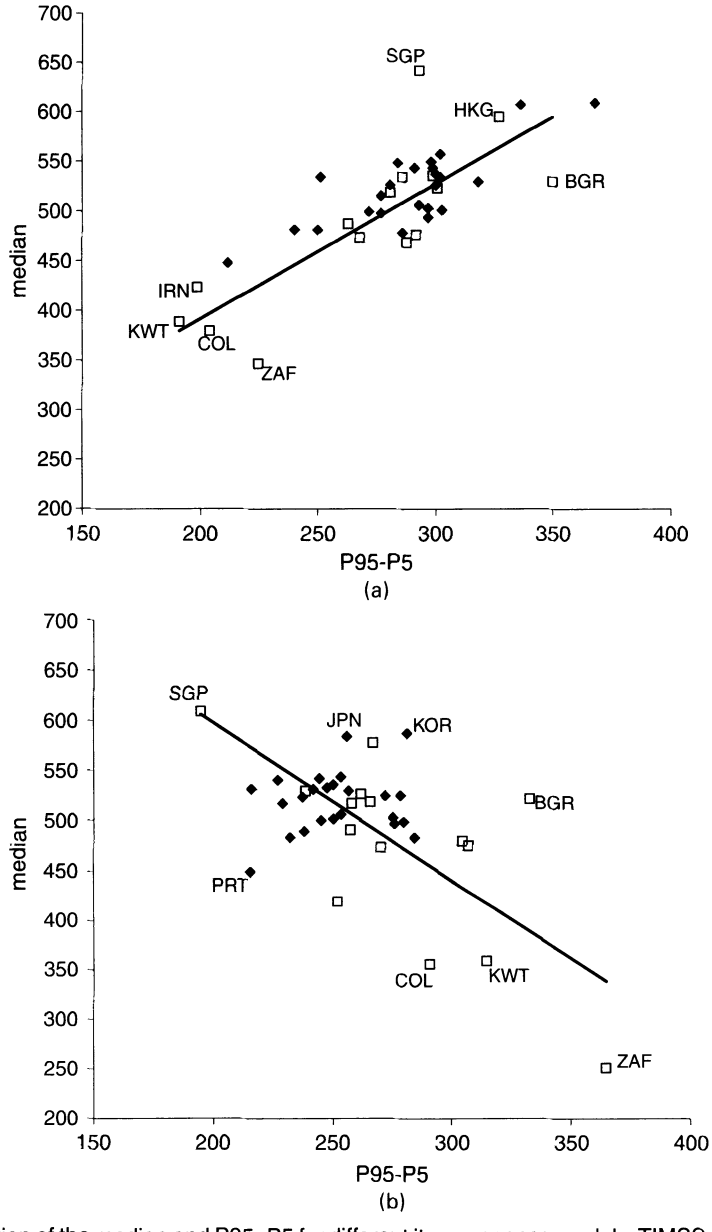

Fig. 7. Association of the median and P95–P5 for different item response models, TIMSS 1995 (the correlations of the median and P95–P5 are 0.79 for the one-parameter (0.78 for OECD countries) and −0.58 for the three-parameter values (0.16 for OECD countries); ♦, OECD countries; □, other countries; SGP, Singapore; HKG, Hong Kong; JPN, Japan; KOR, Korea; BGR, Bulgaria; IRN, Iran; PRT, Portugal; KWT, Kuwait; COL, Columbia; ZAF, South Africa): (a) one-parameter values; (b) three-parameter values

Our hypothesis is that correlations will be higher for the comparisons involving the one-parameter scores since the results are based on the same type of item response model. The hypothesis is rejected—Table 6. The lower correlation for the results that are based on the one-parameter scores is difficult to understand and the size of the change underlines once again that choice of item response model can have major consequences.

Table 6. Correlations of one-parameter and three-parameter values of the median and P95–P5 in the 1995 TIMSS with PISA values

Model	*Results for all 30 countries*		*Results for 23 OECD countries*	
	Median	*P95–P5*	*Median*	*P95–P5*
TIMSS 3 parameter	0.60	0.40	0.70	0.17
TIMSS 1 parameter	0.58	0.14	0.69	0.01

5. Conclusions

There is continued development of international surveys of educational achievement and functional literacy. Users will have increasingly more data available, both in the form of summary statistics in published reports from survey organizers and as microdata available for secondary analysis. It is therefore important that a comparison is made of the surveys' results and analyses are undertaken into the sensitivity of results to the choice of item response model.

We have focused on cross-country patterns of central tendency and dispersion among children and young people aged (depending on the survey) from 10 to 24 years. The broad conclusion from comparing four surveys is that there is a reasonable degree of agreement on both aspects of the national distributions. This is encouraging, although care is needed when assessing the overall record of individual countries. Some countries stand out as performing well. Finland and the Netherlands have high average performance and within-country differences that are smaller than elsewhere. The UK appears on balance as a high dispersion country by OECD standards (although not every survey shows this) as are New Zealand and the USA. Within-country differences tend to be smaller where average achievement is higher.

Our investigation of two item response models that are used by survey organizers shows cross-country patterns of central tendency to be robust to the choice of model. But the same is not true for dispersion, for which model choice can have a big effect. Results on dispersion for less developed countries are much less robust than for OECD countries. This is worrying given the trend over time for the achievement surveys to cover more diverse sets of countries in terms of development level. Even the conclusion over the direction of association between central tendency and dispersion was sensitive to the choice of model when we looked at the group of all countries who participated in the 1995 TIMSS, irrespective of their level of development. We believe that survey reports should include an analysis of the sensitivity of basic results to model choice.

Acknowledgements

This research has in part been supported by a grant from the United Nations Educational, Scientific and Cultural Organisation Institute for Statistics, Montreal. (The views that were expressed are our own and should not be associated with that organization.) We have benefited from discussions with organizers of the TIMSS and PISA surveys (Michael Martin, Ina Mullis, Eugene Gonzalez and Andreas Schleicher) and we are very grateful to them for their help and

comments but they are not responsible for the ways in which we have analysed or represented the data. Useful comments were also made by Stephen Jenkins, Harvey Goldstein and the journal referees and Joint Editor.

Appendix A: Data used in the paper

Details of the surveys are given in their reports: Mullis *et al.* (2000, 2003), Organisation for Economic Co-operation and Development and Statistics Canada (2000), Organisation for Economic Co-operation and Development (2001) and Organisation for Economic Co-operation and Development and United Nations Educational, Scientific and Cultural Organisation Institute for Statistics (2003). The TIMSS and PIRLS surveys are projects of the International Association for the Evaluation of Educational Achievement. The Association's designated study centre for these surveys is the TIMSS and PIRLS International Study Center at Boston College. The OECD secretariat has overall managerial responsibility for the PISA survey.

Besides the eighth-grade children who were analysed in this paper, the 1995 TIMSS collected data which we do not use on children in the third, fourth and seventh grades and children in their final year of secondary schooling. We use TIMSS data on eighth-grade children from 1999 if a country participated in that survey and from 1995 if not (which was the case for Austria, Denmark, French-speaking Belgium, France, Germany, Greece, Iceland, Ireland, Norway, Portugal, Scotland, Spain, Sweden and Switzerland). The 1995 data that are used in Section 3 are those which were derived from a three-parameter item response model and hence provide results on the same basis as those from the 1999 round—see Section 4. (In practice 'eighth grade' in the TIMSS survey means the higher of two adjacent grades in each country that contained the highest proportion of 13-year-old children; the 'fourth grade' in the PIRLS survey means the higher of two adjacent grades that contained the highest proportion of 9-year-old children.) We discuss conditions for direct comparison of three- and one-parameter scores in Brown *et al.* (2005), footnote 14.

Our TIMSS data for the UK refer only to England and Scotland; the data for England are drawn from the 1999 TIMSS and are combined (with appropriate weights to account for differences in population size) with data (three-parameter scores) for Scotland drawn from the 1995 TIMSS. PIRLS data for the UK also refer to England and Scotland only. For the PISA survey, the UK is represented by England, Scotland and Northern Ireland. The IALS covers all parts of the UK. For Belgium, we combine TIMSS 1999 data for Flemish-speaking areas with 1995 data (three-parameter scores) for French-speaking areas. IALS data refer to Flanders only. For Canada, PIRLS coverage is restricted to the provinces of Ontario and Quebec. For Norway, IALS results are restricted to speakers of Bokmal Norwegian, which is the language of the large majority of Norwegians.

In all four surveys, the item response modelling results in five 'plausible values' of proficiency for each individual rather than a single figure. We follow the survey organizers' practice of calculating all summary statistics (e.g. the median or any other percentile) with each plausible value and then averaging the five resulting estimates.

Appendix B: Calculations of average ranks and average *z*-scores

Figs 1 and 2, and Tables 2 and 3 show each country's average ranks and average z-scores for central tendency, measured by the median, and dispersion, measured by the difference between the values of the 95th and fifth percentiles, which we label P95–P5. The calculation of these values may be illustrated with the example of Italy.

Italy's median scores in each of the eight tests that are analysed in Fig. 1 and Table 2 were 492, 462, 480 (PISA reading literacy, mathematics and science respectively), 482, 496 (TIMSS mathematics and science) and 271, 272 and 277 (IALS document, quantitative and prose literacy). These scores placed Italy in the following ranks for the pool of the 18 OECD countries in question: 16, 17, 17, 17, 16, 15, 16 and 15 respectively. The simple average of those ranks is 16.1. However, we weight each survey, PISA, TIMSS and IALS, equally so that the average rank that enters Table 2 for Italy of 16.2 is equal to $\{(16+17+17)/3+(17+16)/2+(15+16+15)/3\}/3$. The z-score for the median for each test is calculated by subtracting from Italy's median the average of the medians for the 18 countries under consideration and then dividing by the standard deviation of these medians; for example, for PISA reading literacy, the value in Table 2 of -1.09 is equal to $(492.4-516.1)/21.8$. Italy's average z-score for the median, -1.19, is calculated in an analogous way to the average rank (i.e. weighting the three surveys equally).

The same methods apply to the calculation of Italy's average rank and z-score for dispersion, measured by P95–P5, shown in Fig. 1 and Table 3. For example, the fifth and 95th percentiles of PISA reading literacy for Italy are equal to 330.9 and 627.5 respectively; hence P95–P5 = 296.6. The values of P95–P5 for Italy are calculated in this way for each test. Italy is then ranked on these values for each test among the pool of 18 OECD countries. The average rank (10.7 for Italy) is calculated in the analogous way as described above for the median (again weighting the three surveys equally). Italy's z-score for P95–P5 for PISA reading of -1.19 is equal to 296.6 minus the average value of P95–P5 for the 18 countries, 323.4, divided by the standard deviation of the 18 P95–P5-values, 22.6. The average z-score for P95–P5 of 0.14 for Italy is the average of the eight z-scores in Table 3, weighting equally the surveys rather than the tests.

References

Adams, R. J. (2002) Scaling PISA cognitive data. In *PISA 2000 Technical Report* (eds R. Adams and M. Wu), pp. 99–108. Paris: Organisation for Economic Co-operation and Development.

Adams, R. J. (2003) Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxf. Rev. Educ.*, **29**, 377–389.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stranat, P., Tillmann, K.-J. and Weiß, M. (eds) (2001) *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im Internationalen Vergleich*. Opladen: Leske–Budrich.

Beaton, A., Mullis, I., Martin, M., Gonzalez, E., Kelly, D. and Smith, T. (1996a) *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill: Boston College.

Beaton, A., Mullis, I., Martin, M., Gonzalez, E., Kelly, D. and Smith, T. (1996b) *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill: Boston College.

Blum, A., Goldstein, H. and Guerin-Pace, F. (2001) An analysis of international comparisons of adult literacy. *Assessmnt Educ.*, **8**, 225–246.

Brown, G., Micklewright, J., Schnepf, S. and Waldmann, R. (2005) Cross-national surveys of learning achievement: how robust are the findings? *Applications and Policy Working Paper A05/05*. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton. (Available from `http://eprints.soton.ac.uk/16250`.)

Brown, M. (1999) Problems of interpreting international comparative data. In *Comparing Standards Internationally: Research and Practice in Mathematics and Beyond* (eds B. Jaworski and D. Phillips), pp. 183–205. Oxford: Symposium Books.

Campbell, J., Kelly, D., Mullis, I., Martin, M. and Sainsbury, M. (2001) *Framework and Specifications for PIRLS Assessment 2001*, 2nd edn. Chestnut Hill: Boston College.

Denny, K. (2002) New methods for comparing literacy across populations: insights from the measurement of poverty. *J. R. Statist. Soc.* A, **165**, 481–493.

Esping-Andersen, G. (2004) Unequal opportunities and the mechanisms of social inheritance. In *Generational Income Mobility in North America and Europe* (ed. M. Corak), pp. 289–314. Cambridge: Cambridge University Press.

Goldstein, H. (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *Br. J. Math. Statist. Psychol.*, **33**, 234–246.

Goldstein, H. (2000) IALS—a commentary on the scaling and data analysis. In *Measuring Adult Literacy: the International Adult Literacy Survey (IALS) in the European Context* (ed. S. Carey). London: Office for National Statistics.

Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA debate. *Assessmnt Educ.*, **11**, 319–330.

Harkness, J., Van deVijver, F. and Mohler, P. (eds) (2002) *Cross-cultural Survey Methods*. Chichester: Wiley.

Mullis, I., Martin, M., Gonzalez, E., Gregory, K., Garden, R., O'Connor, K., Chrostowski, S. and Smith, T. (2000) *TIMSS 1999 International Mathematics Report*. Chestnut Hill: Boston College.

Mullis, I., Martin, M., Gonzales, E. and Kennedy, A. (2003) *PIRLS 2001 International Report*. Chestnut Hill: Boston College.

O'Leary, M. (2001) The effects of age based and grade based sampling on the relative standing of countries in international comparative studies of student achievement. *Br. Educ. Res. J.*, **27**, 187–200.

O'Leary, M. (2002) Stability of country rankings across item formats in the Third International Mathematics and Science Study. *Educ. Measmnt Issues Pract.*, **21**, 27–38.

O'Leary, M., Kellaghan, T., Madaus, G. and Beaton, A. (2000) Consistency of findings across international surveys of mathematics and science achievement: a comparison of IAEP2 and TIMSS. *Educ. Poly Anal. Arch.*, **8**, 43.

Organisation for Economic Co-operation and Development (2001) *Knowledge and Skills for Life—First Results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development and Statistics Canada (2000) *Literacy in the Information Age—Final Report of the International Adult Literacy Survey*. Paris: Organisation for Economic Co-operation and Development.
Organisation for Economic Co-operation and Development and United Nations Educational, Scientific and Cultural Organisation Institute for Statistics (2003) *Literacy Skills for the World of Tomorrow—Further Results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.
Prais, S. J. (1997) Whole-class teaching, school-readiness and pupils' mathematical attainments. *Oxf. Rev. Educ.*, **23**, 275–290.
Prais, S. J. (2003) Cautions on OECD's recent educational survey (PISA). *Oxf. Rev. Educ.*, **29**, 139–163.
Social Exclusion Unit (2001) *Preventing Social Exclusion*. London: Social Exclusion Unit.
United Nations Development Programme (2000) *Human Development Report*. New York: United Nations Development Programme.
Wößmann, L. (2003) Schooling resources, educational institutions and student performance: the international evidence. *Oxf. Bull. Econ. Statist.*, **65**, 117–170.