DISCUSSION PAPER SERIES

IZA DP No. 1652

Cross-National Surveys of Learning Achievement: How Robust are the Findings?

Giorgina Brown
John Micklewright
Sylke V. Schnepf
Robert Waldmann

July 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Cross-National Surveys of Learning Achievement: How Robust are the Findings?

Giorgina Brown
ISTAT, Rome

John Micklewright
S3RI, University of Southampton and IZA Bonn

Sylke V. Schnepf
S3RI, University of Southampton and IZA Bonn

Robert Waldmann
University of Rome Tor Vergata

Discussion Paper No. 1652
July 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1652
July 2005

ABSTRACT

Cross-National Surveys of Learning Achievement: How Robust are the Findings?*

International surveys of learning achievement and functional literacy are increasingly common. We consider two aspects of the robustness of their results. First, we compare results from four surveys: TIMSS, PISA, PIRLS and IALS. This contrasts with the standard approach which is to analyse a single survey with no regard as to whether it agrees or not with other sources. Second, we investigate whether results are sensitive to the choice of item response model used by survey organisers to aggregate respondents' answers. In both cases we focus on countries' average scores, the within-country differences in scores, and on the association between the two. There is mixed news to report.

JEL Classification: I21, J13

Keywords: educational achievement, test scores, IALS, PISA, PIRLS, TIMSS

Corresponding author:

John Micklewright
Southampton Statistical Sciences Research Institute (S3RI)
University of Southampton
Southampton SO17 1BJ
UK
Email: jm4@soton.ac.uk

* This paper has in part been supported by a grant from UNESCO Institute for Statistics, Montreal. (The views expressed are our own and should not be associated with UNESCO.) We have benefited from discussions with organisers of TIMSS and PISA (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.

1. Introduction

Recent years have seen an increasing number of international surveys of learning achievement of children and 'functional' literacy of adults. These include the 1994-8 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). Results from further rounds of all four surveys (or from new successor surveys) will be published in the coming years and the existing data are already used widely across the world by governments and international organisations and by researchers from an increasing range of disciplines.[1]

One feature of all this activity is that the surveys are typically analysed in isolation from one another with no indication as to whether new results confirm or contradict those from earlier surveys. Should we see the maths and science results in PISA 2000 as 'overwriting' those in TIMSS 1995 and 1999, and the reading results in PIRLS 2001 as replacing those from PISA the year before? Obviously this is the wrong approach. Each survey has its merits and defects, and its own particular focus. The age groups studied, the approach to assessment, the form of the tests, and the survey response rates all vary. In short, there is ample reason for comparing results from the different surveys rather than relying on a single source. There have been valuable contributions to this endeavour, usually focused on a few countries, a pair of surveys and one subject (e.g. Prais 1997, 2003). But to our knowledge there has been no study that has pulled together results for all subjects from all the surveys mentioned above for a large group of countries to see whether key dimensions of the pattern of their results are broadly consistent from survey to survey. Making such a comparison is the first contribution of this paper.

Comparing findings from survey to survey 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' organisers using item response models from the psychometric literature. It is a lot more complex than simply adding up the number of correct answers given by each respondent. In contrast to the situation for the more obvious issues listed above, such as age group or approach to assessment, 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

[1] Examples include the UK government's analysis in Social Exclusion Unit (2001), the Human Poverty Index-2 in UNDP (2000), and from disciplines outside educational assessment, Denny (2002) in statistics, Wößmann (2003) in economics, and Esping-Andersen (2004) in sociology.

alternative models have been applied to the data for selected countries, e.g. Blum et al (2001) for IALS and Goldstein (2004) for PISA. But this remains an under-researched area. Our second contribution is to show the extent to which the full cross-national pattern of results from one survey changes with the use of two variants of a standard item response model.

In both parts of the analysis – comparison of survey results and analysis of sensitivity to scaling method – we focus on two substantive issues. The first is the cross-country pattern of central tendency and of dispersion. Both aspects of the national score distributions are of obvious interest. How well children and young people in any country are doing on average in absolute terms is important to know in a globalised world. But we also need to see the extent of educational inequalities within each country, inequalities that can be expected to help 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. Are different surveys in broad agreement on these two features of the national distributions and does the scaling method change the picture? And are there particular countries that are consistent outliers, with for example high dispersion? For example, it has frequently been suggested that the UK is one such country. We consider the evidence across the different surveys for this.

The second issue is the relationship of central tendency to dispersion. This is also a topic of natural interest for educational policy. Do the different surveys and different scaling methods provide a clear picture of the association of these two basic features of score distributions? Do they show for example that countries with higher mean achievement have lower dispersion, or vice versa, or is there no apparent association between the two?

In focusing on these basic issues concerning the position and shape of the national distributions we neglect other aspects of the surveys' results. In particular, we do not consider whether correlates of achievement vary across surveys, or with the scaling method. For example, it would be of considerable interest to see whether a robust picture exists of the relationship of scores with socio-economic background or with characteristics of schools.[2]

Section 2 provides a brief description of the four surveys we consider, focusing on features that differ among them. Section 3 shows how results for central tendency and dispersion vary across these four sources of achievement data. Section 4 reports on our analysis of robustness of results to choice of item response model. We concentrate on TIMSS 1995 where results based on two different item response models are available from the survey organisers (we also discuss implications for comparisons across surveys). Section 5 concludes.

[2] Some results on correlates, with a focus on family structure, are given in Micklewright and Schnepf (2004, 2005). Research in this area is limited by differences between surveys in the variables that are collected.

2. The international achievement surveys

Table 1 lists the data we use from the different sources. Three surveys – PIRLS, TIMSS and PISA – collect data on children of school age. Their designs involve the selection of a sample of schools and then a randomly selected single class (TIMSS and PIRLS) or sample of all pupils (PISA) within each school. Sample size averages about 4,000 to 6,000 children per country, depending on the survey. By contrast, the fourth survey, IALS, is a household survey that collected information on all people of working age; we restrict attention to the young people aged 16-24, of which there are on average about 700 per country.[3] In the case of TIMSS, we use data from both the 1995 and 1999 rounds, taking the earlier year if a country did not participate in the later round.[4]

The first practical issue confronting comparisons of their results is that country coverage varies from survey to survey. However, there are reasonable sized groups of countries that are covered by at least three surveys. In Section 3 we concentrate on a group of 18 countries that are present in TIMSS, PISA and IALS and a group of 21 in TIMSS, PISA and PIRLS. (Only 11 countries are in all four surveys).[5] The first group has the merit of being composed entirely of OECD members, i.e. countries at broadly similar (and high) levels of national income. This means that cross-country differences are not driven by factors associated with large differences in levels of development. By contrast, the second group contains 14 OECD members (of which 11 are also in the first group) and seven other countries: two rich ones, Hong Kong and Israel, and five Central and Eastern European countries at significantly lower levels of development: Russia, Latvia, Bulgaria, Macedonia, and Romania. In Section 4 we use all 38 countries in TIMSS 1995 for which microdata are available, of which only 24 are from the OECD and 11 are classified by the World Bank as middle or low income. The distinction between rich and poor countries turns out to be important when considering the sensitivity of results to choice of item response model. (Further details of our selection of data are given in the Appendix.)

The four surveys collect information on performance in a number of subjects. TIMSS and PISA both cover achievement in maths and science. PISA in addition measures reading

[3] Details on the surveys can be found in their reports: Mullis et al (2000), Mullis et al (2003), OECD and Statistics Canada (2000), OECD (2001) and OECD and UNESCO Institute for Statistics (2003). TIMSS and PIRLS are projects of the International Association for the Evaluation of Educational Achievement (IEA). The IEA's designated study centre for these studies is TIMSS and PIRLS International Study Center at Boston College. The OECD Secretariat has overall managerial responsibility for PISA.

[4] TIMSS 1995 also covered 3-4th graders and children in their final year of secondary schooling. We do not use those data in this paper. The 1995 TIMSS data that we use in Section 3 are those derived from a 'three-parameter' item response model and hence provide results on the same basis as the 1999 round – see Section 4 for more details.

[5] Statements about country coverage refer to the survey rounds shown in Table 1.

ability, which is the (sole) focus of PIRLS. IALS measures 'quantitative', 'prose' and 'document' literacy. The first of these uses a mathematical skill (essentially arithmetic) while the second requires reading skills. For some of our analysis we combine information for each country across different subjects so as to get a broad summary view of national performance. However, we also look at individual subjects, for example how countries' TIMSS maths scores compare to their PISA maths scores (considering separately central tendency and dispersion). But our goal of looking at the big picture across several surveys means that in contrast to some authors we do not disaggregate into different aspects of each subject within each survey. For example, we do not consider how the UK performs in geometry relative to other countries as opposed to other aspects of maths (e.g. see discussion by Brown 1999 who used TIMSS data). For convenience we refer to all four surveys as measuring 'achievement' in the subjects that they cover and to each subject in a survey as a 'test'. Hence for the group of 18 countries in TIMSS, PISA and IALS we have information on achievement in eight tests while for the 21 country group in TIMSS, PISA and PIRLS there is information from six tests.

Country and subject coverage aside, the surveys differ in a number of obvious dimensions that can be expected to contribute to differences between them in results: type of achievement assessed and the degree of detail with which this is done, age of target population, levels of response, the form of test, and the item response model applied to the data.

Type of achievement. If PISA were to measure ability in maths in exactly the same way as TIMSS there would be little rationale for both surveys assessing this subject. Likewise, if their reading assessments were identical, the value-added of PIRLS over PISA would be just the focus on a younger age group. However, there are differences across the surveys in the types of achievement that are assessed. One should not expect them to give identical cross-country patterns of results for the apparently same subject, e.g. 'maths'. But if results were to differ sharply one would worry that the particular choice of measuring rod is driving each survey's results.

IALS tries to measure the extent to which people can use literacy skills of different types to perform everyday tasks. For example, respondents are asked to work out a tip, determine the amount of interest on a loan, and extract appropriate information from a transport timetable. PISA also has a focus on use of knowledge to address problems that arise in real-life settings and similarities in conceptual approach to that of IALS are emphasised in

the first report (OECD 2001: 18).[6] By contrast, TIMSS measures mastery of internationally agreed curricula and there is variation in the extent to which these match any individual country's actual national curriculum in maths or science. While these differences between the foci of PISA and TIMSS are well known, it is less clear how PISA and PIRLS differ in their approach to reading. The PIRLS organisers argue that the approaches are similar, with both being based on 'an expanded notion of literacy' (Campbell et al 2001: 85), and that differences are due principally to the need to take account of the different age-groups being assessed.

Detail in assessment. Different surveys may cover the same subject area in differing degrees of detail. TIMSS and PISA both assess maths and science. But the TIMSS 1999 maths and science assessments had 169 and 146 items respectively, compared to the 32 items measuring literacy in maths in PISA 2000 and the 35 in science. PISA in that year concentrated on reading, with the assessment of maths and science taking second place. (PISA 2003, not used in this paper, concentrates on maths and PISA 2006 will concentrate on science.) By contrast the comparison of the number of items to assess reading in PISA 2000 and PIRLS 2001 is more balanced, with 98 items in PIRLS and 141 in PISA.

Age group. Most of the differences across the surveys are obvious. PIRLS covers young children. Our IALS results relate to young people in their late teens and early 20s. The children in PISA and TIMSS are in their early or mid teens. Countries may do well at one age and not at another so that a focus on the results from just one survey may mislead on how well children and young people of all ages perform. One difference across surveys in age coverage is less obvious. PISA measures achievement of children of a given age, while TIMSS and PIRLS cover children in a school 'grade'.[7] Some countries promote all children at the end of the year to the next grade irrespective of their achievement, while others insist on a certain competence being reached before passage upwards is allowed. Where the latter practice is followed, average achievement relative to other countries can be expected to be higher in TIMSS than in PISA. But the same countries might show higher disparities in achievement in PISA.

Response. The surveys differ in their response rates. Among the 21 countries in TIMSS, PISA and PIRLS, the overall rates of response (taking into account both non-response at school and student level) averaged 83 percent for PISA, 89 percent for TIMSS

[6] However, one of the PISA organisers has stressed that, in the case of maths, the survey's approach is not the same as assessing 'everyday functional mathematical literacy' (Adams 2003: 379).

[7] TIMSS 1995 assessed children in the two adjacent grades that contained the highest proportion of 13 year olds, which were typically 7th and 8th grades. TIMSS 1999 tested children in the upper of these two grades only. PIRLS assessed children in the upper of the two grades with the most 9 year olds at the time of testing. This corresponds to the 4th grade and an average age of about 10 years for most countries.

and 90 percent for PIRLS. Response to IALS in all participating countries averaged 63 percent but this refers to adults of all ages and not just to the 16-24 year olds focused on in this paper. The correlation in the country response rates between surveys is certainly positive, but not that high: 0.51 for PISA-TIMSS, 0.38 for PISA-PIRLS and 0.42 for TIMSS-PIRLS. It is unlikely that any non-response bias affecting estimates of central tendency or dispersion of scores for any country would be the same for all surveys. Non-response seems to be a particular problem in the UK. The overall response rate for England was 16 points below the average in PISA, 12 points below in TIMSS and 8 points below in PIRLS. Investigation of possible bias for England concluded that the impact was 'negligible' (OECD 2001: 236) although analysis appears to have been limited to school-level non-response and to have considered only central tendency. Concern about non-response bias led to results for the UK being excluded from the international report for PISA 2003 (data not analysed in this paper).

Form of testing. About two-thirds of the TIMSS questions in 1999 were multiple choice, significantly more than in PISA. Only about a third of the PIRLS assessment (in terms of possible scores) is based on multiple-choice. IALS has no multiple choice element. These differences may affect countries' relative standings in the different surveys: arguably children in some countries do better at multiple choice questions than children in others due to variation in countries' traditions of multiple-choice testing in schools.

Finally, there are differences in how answers to questions are aggregated. A respondent's answers are summarised into a single score for the subject concerned – maths, science, reading, different types of literacy, etc.[8] We defer further 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. In each survey, scores are scaled to produce values that are chosen by the survey organisers for the mean and standard deviation among all persons in participating countries – 500 and 100 respectively in TIMSS, PISA and PIRLS, and about 275 and 20 in IALS. However, 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. TIMSS and PIRLS both include a wider range of countries in terms of level of development than does PISA, which is largely focused on OECD members. So, for example, that Italy had a mean reading score of 541 in PIRLS but only 487 in PISA may in part merely reflect the fact that PIRLS included such countries as Belize, Columbia and Morocco whereas the PISA scale is based solely on the OECD countries participating in the 2000 round.

[8] In fact the item response modelling results in five 'plausible values' for each individual rather than a single figure. We follow survey organisers' 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.

We use two methods to overcome this problem. First, within each of the two groups of countries present in three surveys, we compare the *rankings* of countries 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, but we stress that our use of rankings is not intended to propagate that view – we rank in order to compare more easily across tests. Second, we convert the measures of central tendency and dispersion for each country into *z-scores*, a standard method for obtaining comparable units of measurement. That is, for the pool of 18 countries covered by PISA, TIMSS and IALS and the 21 in PISA, TIMSS and PIRLS, 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. (Country rankings on z-scores are the same as on the untransformed scores and the correlations between the country values are also unchanged by the z-score transformation.)

3. Comparing results across surveys

Our objective in this section is to assess whether different surveys and subjects give a similar picture of the cross-country pattern of central tendency and dispersion. We measure central tendency by the median and dispersion by the difference between 5th and 95th percentiles, P95-P5, using z-score transformations in the way just described.[9]

We start by giving a graphical summary that includes all eight tests in PISA, TIMSS and IALS for the 18 countries covered by all three surveys. Figure 1 plots each country's *average rank* on the median and on P95-P5 against each other, weighting the surveys equally (rather than the tests). Six countries are identified by a different symbol for a reason explained below. The average ranks have considerable merit as quick 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 likely 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. 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.)

[9] These choices are not important to the results.

Three features of the patterns in Figure 1 stand out. First, the average ranks, both for the median and for P95-P5, display considerable variation. Our first substantive question outlined in the Introduction was whether the different surveys give a similar cross-country picture of central tendency and dispersion. The variation in average ranks is encouraging evidence in favour of 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 in general have smaller within-country differences. This starts to answer our second substantive question which is on the relationship between central tendency and dispersion. Note that the pool of countries being considered here is drawn exclusively from the OECD area so the pattern in Figure 1 is not a result that is produced by mixing countries at very different levels of development.

Third, a number of countries are obvious outliers. Finland has an average rank of only 3.7 on the median and 2.4 on P95-P5. At the opposite end of the spectrum the USA averages 13.6 and 16.7 respectively on the two measures. The countries identified by a different symbol are the six English-speaking ones and two others of these, the UK and New Zealand, join the USA in having the largest within-country differences. Although the achievement surveys have instruments that are designed to work equally well in any language, inevitably there are concerns that full comparability is not obtained. The common language of these six countries removes the problem as far as comparisons between them are concerned. It is notable that there are some substantial differences within this group with Canada and Australia standing in marked contrast to their nearest English-speaking neighbours. Among the full group of 18 countries, Italy and Portugal stand out as exceptions to the general pattern of association of central tendency and dispersion. Despite only mid-table positions on dispersion (in average rank terms) they have very high average ranks 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 columns 3-10 indicates the third of the distribution for that test in which a country falls: dark shading for lowest third, light shading for middle third, white for top third. The countries are ordered on the basis of the average ranks used in Figure 1. The values of these averages are given in column 1 and the average z-scores (again weighting surveys equally) are given in column 2.

The mean of the country average z-scores is by definition equal to zero (we comment in the appendix on the standard deviation).[10]

Table 2 shows both Finland and the Netherlands to have medians that on average are more than one standard deviation above the group mean. In no test are they ever below the mean although both fall into the middle third of the distribution in one case. Portugal, at the other extreme, averages two standard deviations below the mean, which is far more than the only other country with an average z-score below minus one, Italy. In the middle of the distribution, the UK's average rank of 10.1 reflects a considerable mix of results for individual tests: there are two white, three light grey and three dark grey cells, with z-scores varying from 1.12 for PISA science to -0.93 for IALS quantitative literacy. And while all the UK's PISA z-scores are positive, all those for 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 and a half of all 18 countries have three colours in their row of entries.

The same pattern is found in Table 3: a half of the countries have three colours. But the UK is not among them this time. The UK's lower average rank in Figure 1 for the within-country differences reflects greater agreement among the different tests, resulting in P95-P5 values that average 0.8 standard deviations above the group mean. However, it is notable that there is some disagreement for the UK between PISA on the one hand and TIMSS and IALS on the other. In the case of PISA compared to IALS, this repeats the pattern in Table 2: the UK performs better in PISA. Germany is an interesting case of disagreement between PISA and the other two surveys that runs the other way: the large within-country differences for Germany in PISA have been much commented on (e.g. Baumert et al 2001) whereas IALS gives a quite different picture: score dispersion for 16-24 year-olds Germans in this survey is among the smallest for the 18 countries. Moving to the top of the table, Finland's very low average rank of 2.4 is reflected in white cell entries across the board.

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

The new countries of Macedonia, Romania and Israel stand out as having low average achievement and high dispersion. Macedonia repeats Portugal's position in Figure 1: bottom

[10] Average ranks and average z-scores have a correlation of 0.95 for the median and -0.98 for P95-P5.

rank on the median in every test. Romania and Israel, along with Greece, are in the bottom third of the distribution for all tests for the median. Hong Kong on the other hand is found firmly in the opposite quadrant and with the best average rank on within-country differences of any country. These are clear results, both for the countries concerned and in terms of re-enforcing the pattern of association between central tendency and dispersion in Figure 1: on average within-country differences are lowest where average scores are highest. However, even for an apparently clear case like Hong Kong, there is a need to keep an eye on more than one survey. Hong Kong's average rank of 6 on the median (and a z-score average of -0.99) includes a PIRLS rank of 13 (and z-score of -0.07), right out of line with the results from TIMSS and PISA.

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 IALS, in which the UK performed badly, with PIRLS where the UK did well (ranking 3rd with a z-score of 0.89). However, on dispersion the UK once again stands out in PIRLS as a country with high within-country differences (P95-P5 z-score of 0.99). The situation is similar for USA and New Zealand: their relative positions improve on both median and P95-P5 due to the change in the country pool but the substitution of PIRLS for 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 Figures 1 and 2 does not change the conclusion that these three countries have large within country differences by international standards.

One disadvantage of the average rank and z-score calculations is that they give an equal weight 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 PISA and IALS.) The former can be expected to be stronger, and one might well want to take less notice of it. It is the agreement between *surveys* rather than between tests that we may want to focus on (as reflected already in some of our discussion of Tables 2 and 3).

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 21 country groups. Are the correlations within survey for different subjects higher than those between surveys for similar subjects? In Table 4, the answer is 'yes': 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 only exception is for the TIMSS medians where the correlation between

maths and science is less than that between TIMSS science and PISA science (0.66 compared to 0.72). The same pattern is found in Table 5 where the inclusion of countries at lower levels of development pushes up the within survey correlations of country scores in PISA and TIMSS.

On the other hand, it is also true that among the correlations between tests from different surveys, it is typically the case that the values for subjects that are similar are *higher* than those for other subjects (although there are exceptions). For example, the correlation is 0.72 between the science medians in TIMSS and PISA in Table 4 and 0.65 for the maths medians, higher than other combinations of TIMSS and PISA tests. This seems encouraging for our confidence in the general message to be obtained about a subject from each survey.

A general point to note from comparing Tables 4 and 5 is that the correlations for P95-P5 are typically lower than for the median. The average (off-diagonal) correlation coefficients in Table 4 for the two are 0.48 and 0.56 respectively and in Table 5 are 0.55 and 0.69. There is more agreement between tests on the country pattern of central tendency than for dispersion. That does not seem surprising, the latter being harder to measure well. And as we will see in Section 4, the measurement of dispersion appears to be much more sensitive to choice of item response model, which may differ from survey to survey.

We undertook two sensitivity analyses for the correlations between tests. The first concerns the age of respondents. Correlations between test results in TIMSS and PISA might be expected to be higher (ceteris paribus) than those between either survey and PIRLS or IALS on account of the similarity in the ages of children covered. However, PISA surveys children of a given age while TIMSS targets a school grade. Section 2 noted possible consequences for comparison of results from the two sources. To try to adjust for the difference in approach, we re-calculate PISA-TIMSS correlations using sub-samples of children of the same age from TIMSS and of the same grade from PISA.[11] Within the pool of 21 countries covered by PISA, TIMSS and PIRLS the effect is to raise PISA-TIMSS correlations for the median only slightly (by 0.04 for maths and by 0.02 for science) and those for P95-P5 rather more substantially: by 0.07 for maths and by 0.12 for science. For the 18 country pool focused on in Table 4, the PISA-TIMSS correlations for P95-P5 increase by 0.07 for science and as much as 0.22 for maths. However, for the median the correlations *fall*: by 0.07 for both these subject pairs. Hence within-country differences in the two surveys do appear to become somewhat more similar when the age-grade selection is made more homogenous but there is ambiguity for central tendency.

[11] We take children born 14 years before the test year in TIMSS and those two grades above the TIMSS grade in PISA. On average, the sub-samples represent 60 percent of the full TIMSS 8th grade samples and 62 percent of the full PISA samples (for the pool of 21 countries).

The second issue is the effect of sampling error, which is one reason why we see disagreement between tests from different surveys. Even if the four surveys were to be made identical in every aspect of design (target population, sampling scheme, survey instrument etc), sampling error would mean that they would still not produce results that correlate perfectly with each other since their results would be based on different samples of individuals. In practice sampling error is much more of an issue for P95-P5 than for the median. We use available information on standard errors in TIMSS, PISA and PIRLS to estimate the impact of sampling error on the correlations.[12] For the median, we estimate that the correlations in Table 5 would typically increase only by 0.01-0.02 if sampling error were eliminated completely. However, allowing for sampling error pushes up the correlations for tests in different surveys for P95-P5 by an average of 0.07, which is not unappreciable. This is sufficient to close much of the gap between the average (off-diagonal) levels of correlation for central tendency and dispersion that we observed earlier for Tables 4 and 5.

Three broad conclusions come from the comparisons in this section. First, there is considerable agreement on both central tendency and dispersion between the different tests contained in the four surveys, as summarised by average ranks and z-scores. This agreement is sufficient to establish (a) some clear outlier countries for both average achievement and within country differences, and (b) 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, the outlier countries apart, care is often needed in judging the record of individual countries, with the different subjects and surveys not infrequently giving rather different results. Third, agreement between tests in different surveys tends to be less than agreement between tests within the same survey. Amongst other things, this underlines the importance of considering factors that may be peculiar to each survey. These include the item response modelling, the subject of the next section.

4. Comparing item response models

The data analysed in this paper are obtained from item response (IR) models. These models are used by the survey organisers to produce summary scores for each individual and hence to

[12] Standard errors are published for the median, P5 and P95. (We calculate the standard error for P95-P5 following the procedure outlined in Brown and Micklewright 2004.) Consider the case of the median. The sample variance across counties in median achievement is equal to the sum of the true variance and the variance of the sampling error. The latter is given by the sum of the reported standard errors divided by n-1 where n is the number of countries. Hence we can recover an estimate of the true variance of each test median across countries and can re-compute the denominator of the correlation coefficients in Tables 4 and 5 (the correlations for untransformed scores and z-scores are identical). The sample covariance in the numerator is left unadjusted since it is an unbiased estimate of the population covariance.

arrive at estimates of central tendency and dispersion for each country. The achievement scores are therefore *derived* data and the question arises as to whether the choices made over the method of derivation have an appreciable impact on the surveys' results.

Too little is known about this. Typically nothing is said on the subject in the official survey reports. Many users of the surveys access only those published sources. But even where secondary analysis is made of the microdata, the procedures involved in fitting the relevant IR models are sufficiently complex that it is impractical for most researchers to estimate variants so as to gauge the sensitivity of their results to alternative methods.

The patterns of central tendency and dispersion considered in Section 3 might differ from survey to survey simply because survey organisers use different methods for producing summary scores. Or similarities in results may be masking very different modelling of the data. Certain countries might appear as outliers in one survey but not in another because of the particular item response models that are used. And the direction or strength of association between central tendency and dispersion might even hinge on the modelling method. Or the modelling may be irrelevant and we can all sleep comfortably.

This section explores these issues by seeing how results concerning central tendency, dispersion, and the association between the two change for one survey, TIMSS 1995, when two different IR models are applied to the data. This isolates the impact of the IR model on the results, which cannot be seen when simply comparing results from different surveys. We then comment on what this may imply for differences in results across surveys given what is known about the different types of IR model that each survey's organisers have used.

The IR models employed by the survey organisers are invariably 'unidimensional', appropriate when high ability individuals have a greater probability than low ability individuals of answering each and every question correctly. Goldstein (2000, 2004) has criticised this aspect of the modelling process and has experimented with 'multidimensional' models that allow for the situation where one group of individuals is more able to answer one sort of question correctly but not other sorts. In this paper we restrict attention to unidimensional models, and this can be seen as complementing experiments with a multidimensional approach. Both are concerned with sensitivity of results to the modelling choices.

Within the unidimensional class of models, those applied by the survey organisers are typically 'one parameter' or 'three parameter' varieties. The purpose in both cases is to estimate a person's 'proficiency' in the subject concerned (maths, science, etc) from answers to a number of questions. The one parameter model allows for differences in the degree of difficulty of each question. The three parameter model allows in addition for the probability

that the answer is simply guessed and for the ability of a question to discriminate between students of high and low ability. Formally, the models give the probability of a correct answer to question i by student j as:

One parameter model: $p_{ij}(\text{correct answer}) = 1/[1+\exp(-(\theta_j - \alpha_i))]$

Three parameter model: $p_{ij}(\text{correct answer}) = \gamma_i + (1-\gamma_i)/[1+\exp(-\beta_i(\theta_j - \alpha_i))]$

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 one of these logit models, in which the θ_j are treated as unobserved fixed effects in order estimate the other parameters, is only the first step in the procedure to derive the achievement scores.

Results from TIMSS 1995 have been produced by the survey's organisers with both types of model. A one parameter model was used to obtain the results published in the original survey reports (Beaton et al 1996, 1996a). But results for TIMSS 1999 were produced with the more sophisticated three parameter model, which was then also applied to the 1995 data in order to allow results to be compared over time for countries present in the survey in both years.[13] No systematic analysis appears to have been published of differences in results obtained with the data derived from the two models.[14] However, a full set of 1995 microdata using the three parameter model has been made available for each country on the TIMSS website alongside the one parameter model data (including for those countries not present in the 1999 round of the survey). These two sets of microdata are the basis for the analysis in this section and are available for 38 of the 40 countries covered by TIMSS 1995 (the exceptions are Bulgaria and 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 'three parameter scores', θ was modelled as a linear function of

[13] Our analysis in Section 3 uses TIMSS data for some countries from 1999 and some from 1995 (if a country is not present in 1995); see the Appendix. Where 1995 data are used, the results are from the 1999 model, i.e. 'three parameter scores'.

[14] The technical report for TIMSS 1999 argues that direct comparison of scores from the two models is not appropriate because the three parameter model was estimated with 8th grade students only (the 7th grade was not covered in the 1999 round) whereas the original one parameter model was estimated with both 7th and 8th grade students (Yamamoto and Kulick 2000: 253). This should imply that mean scores for 8th graders in the 1995 data from the three parameter model are slightly lower than the original scores and the variances are slightly higher. However, this difference does not invalidate comparison of the overall cross-country patterns of central tendency and dispersion in the two sets of results of the sort made in this paper.

observable characteristics of the student and his or her school.[15] (The number of conditioning variables for each country is given in the TIMSS 1999 Technical Report but the variables themselves are not listed).

Figure 3 shows the distributions of the two sets of maths scores for four countries, selected to reflect the range of differences that occur. In the case of the UK, the switch in IR model leads to a loss of positive skew but overall the distributions seem similar. The picture is not the same in the other three cases. For Singapore, there is a substantial reduction in dispersion. By contrast, in Iran there is a widening of the distribution, while in South Africa there is both a large reduction in the mean and a large increase in dispersion, together with a move to a much more positively skewed distribution. We surmise that the big 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 as one explanation of correct answers. 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. (There is even a small rise evident in Figure 3 at the top of the South African distribution.)

One difference between the two sets of scores is that the correlation between the derived scores produced from the IR model and the 'raw' scores is lower for the three-parameter model. (This is found in *every* country and not just the four in Figure 3.) These raw scores are simply the initial points each child scored on the test, standardised only within each country for differences in the particular test booklet that a child answered. The average country correlation between the final and raw scores was 0.97 and 0.93 respectively with the one- and three-parameter models for maths and 0.95 and 0.90 for science. Not surprisingly the more complex model pushes the scores further away from the raw data. Of course, this is a good thing if the resulting distribution of achievement better represents reality. However, the extent of the change varies from country to country with some marked outliers. These include Iran, where the correlation for maths is 0.88 with the three-parameter model, Columbia (0.82), South Africa (0.80) and Kuwait (0.79). All four countries have correlations in the range 0.94-0.96 with the one parameter model. Whether the figures for the outliers are low in absolute terms is a matter of judgement – a figure of only 60-65 percent for the explanation of derived scores by raw scores (i.e. for r^2) might seem rather lower than one would hope for.

[15] The maximum likelihood estimate of this model implies a probability distribution for each student's θ. This distribution is used as a prior and updated with the student's actual performance. Then five 'plausible values' for each student are picked from the resulting posterior (see footnote 8).

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 first with central tendency, as measured by the median. Figure 4 plots each country's median for the maths three-parameter scores against that for the one-parameter scores. OECD countries are distinguished by a different symbol. To be clear: the raw data behind the two sets of scores – the answers given by respondents to the questions – are identical. What differs is the method used to summarise those data for each individual into a single number.

The conclusion seems straightforward. The medians are very highly correlated, both among just the OECD countries present in TIMSS 1995 and among all countries covered by the survey in that year. And this is true for both maths and science. The rankings hardly change at all. The cross country pattern of central tendency is robust to the change in IR model. However, for both subjects a few countries lie some way off the 45 degree line. South Africa (ZAF) is the most extreme case for both subjects. There is a fall in the maths median from the one to the three-parameter scores of over 75 points (also shown clearly in Figure 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. The one-parameter median is 2.6 standard deviations below the average for other countries but this increases to 3.4 with the three-parameter data.

We now turn to dispersion, measured as in Section 3 by the difference between 95th and 5th percentiles. As a preliminary, Figure 5 shows what happens to P5 and P95 separately, focusing on maths. In both cases the correlation between one- and three-parameter scores is again very high, as for the median. For P5, countries with low values with the one-parameter model are pushed even lower with the three-parameter model, as happens with the median. Kuwait and South Africa change by over 100 points, i.e. more than for the median.

For P95, the opposite pattern is found. For most countries the value falls somewhat but countries with high values on the one parameter scores have the largest reduction – Japan, Korea and Singapore each see falls of 60-80 points. While those countries with the lowest values – South Africa, Kuwait and Colombia – actually see a slight rise (visible in the comparison of the upper tails in the case of South Africa in Figure 3). A similar pattern of results for both P5 and P95 is found for science.

This difference in the pattern of change for the two quantiles is critical. For P5 the slope of the regression line would clearly be greater than one whereas for P95 it would be less than one. 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.[16]

The net result in terms of change in P95-P5 is shown in Figure 6 for both maths and science. For maths, the correlation between the two sets of values is essentially zero (0.03): in contrast to the median, the cross-country pattern of dispersion is therefore far from robust to the choice of IR model.[17] The change in the position of South Africa is dramatic. 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 and Colombia are almost as striking. The change in dispersion relative to central tendency is even larger for these three countries (given that the median falls for all of them in the switch to the three parameter model). The coefficient of variation for South Africa more than doubles from 0.18 to 0.39. Singapore on the other hand changes from being a middle-ranking country for dispersion of one-parameter scores to being the country with the smallest within-country differences in three-parameter scores.

The absence of any correlation at all for maths is clearly driven in large measure 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 has an average value of P95-P5 for the OECD for the one parameter scores but the greatest (just) dispersion in the OECD group for the three parameter scores. (Since Greece lies on the 45 degree line, this change results from alterations in the values of other countries.)

The situation for science is different. Here the change in IR model has much less impact on the cross-country pattern of dispersion. Nevertheless, there is still some notable re-ranking. For example, Kuwait and Colombia are above the 45 degree line, as for maths. Dispersion increases for them with the three-parameter model and is well above that in Singapore, having been some 75 points below Singapore with the original one parameter model. In the three parameter data, dispersion in the UK and Cyprus is almost identical and the two countries have adjacent ranks. In the one parameter data they are separated by 20 ranks and about 60 points. South Africa becomes a big outlier with much larger dispersion for the three-parameter data, having been merely one of the countries with more dispersed sets of one-parameter scores (in contrast to the situation for maths).

[16] They are in fact 1.58 and 0.68.

[17] If the standard deviation is used in place of P95-P5 the correlations are 0.07 (all countries) and 0.74 (OECD only) for maths, and 0.63 and 0.86 for science: our results are not sensitive to the choice of measure of dispersion.

Finally, Figure 7 shows how the switch in IR model changes the view of whether dispersion rises or falls with central tendency, another of the core issues of the paper, focusing on maths. With the one-parameter data, the conclusion is that countries with higher average achievement have higher dispersion in achievement ($r = +0.79$). With the three-parameter data the opposite conclusion would be drawn: as average achievement rises, inequality in achievement falls ($r = -0.58$). This latter result was one of our conclusions from the comparisons of different surveys in Section 3 (where in the case of TIMSS we used three-parameter data) although the focus there was mainly on the 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 we do not show) are again less dramatic, weak positive correlation switching to weak negative correlation.

To summarise: (i) the country pattern of central tendency in TIMSS 1995 is not sensitive to the choice of one or three parameter model, with the one major exception of South Africa (but this made no difference to the ranking); (ii) the pattern of dispersion for maths is really quite sensitive with some very sharp changes in rankings for some countries that alter completely the picture of the outliers, but there is less sensitivity for the OECD countries and results for science also display much less change; and (iii) the direction of association of central tendency and dispersion for maths changes with the switch in IR model. The greater sensitivity of the results for less developed countries, most notably South Africa, 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 TIMSS.

What do these findings imply for those obtained earlier from comparing different surveys? The TIMSS results in Section 3 are all based on the three-parameter scores. But what is the IR model behind the results for PISA, IALS and PIRLS? Unless it is the same as that for the TIMSS data we have not been comparing like with like.

The models used in IALS and PIRLS are similar to that used to derive the three parameter TIMSS data: comparisons between any of these sources can rely on a high degree of comparability of IR model.[18] However, PISA used a one parameter model, as originally employed for TIMSS 1995.[19] As a consequence, the results in Section 3 for PISA are not from the same type of IR model as those from the other surveys. Our findings in the present section

[18] A two parameter model was used in IALS allowing for difficulty of questions and their ability to discriminate between persons of differing abilities (Yamamoto 1998). No guessing parameter was needed since all questions required 'constructed responses'. PIRLS used two parameter models for questions of this type and three parameter models, allowing in addition for guessing, for multiple choice questions (Gonzalez 2003).

[19] The model is described in Adams (2002) and the procedures are reported elsewhere as being 'identical to those that were used in TIMSS 1995' (Adams 2003: 386).

suggest 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 choice of IR model suggests that comparisons of within-country differences in PISA with those in the other surveys may be potentially misleading.

To explore this issue we take maths score data for countries in both TIMSS 1995 and PISA and compare correlations of central tendency (measured by the median) and dispersion (measured by P95-P5) between (i) TIMSS three-parameter results and PISA results and (ii) TIMSS one-parameter results and PISA results. We consider both all 29 countries in both surveys and the 23 from the OECD. Our hypothesis is that correlations will be higher for the comparisons involving the one-parameter scores since the results are based on the same form of IR model. The hypothesis is rejected – see Table 6. For example, the correlation between P95-P5 values for TIMSS three-parameter data and PISA data is 0.43 but this falls to 0.18 when the TIMSS one-parameter data are used. The lower correlation with the one-parameter data is difficult to understand but whatever the direction of the change its size underlines once again that choice of IR model does not seem to be a trivial issue.

5. Conclusions

The rest of this decade will see continued development of international surveys of learning achievement and functional literacy. Users will have more and more data available to them, both in the form of summary statistics and analyses in published reports from the survey organisers and as microdata sets available for secondary analysis. In this situation it is important that comparison is made of the surveys' results and analysis is undertaken into the sensitivity of results to choice of IR model.

Our focus has been on basic cross-country patterns of central tendency and dispersion among children and young people aged (depending on the survey) from 10 to 24. The broad conclusion from comparing four surveys is that there is a reasonable degree of agreement between the sources available to date on both aspects of the national distributions. This is certainly encouraging, although care is needed when assessing the overall record of individual countries. Some countries do stand out as performing well in all surveys. Finland and the Netherlands have high average performance and within-country differences that are smaller than elsewhere. The UK on the other hand appears on balance as a high dispersion country by OECD standards (although not every survey shows this) as are New Zealand and the USA. (On central tendency, however, the UK does not seem to be towards either extreme.) Finally,

the general cross-country picture given by the surveys taken together is that within-country differences tend to be smaller where average achievement is higher.

Our investigation of two IR models that have been used by survey organisers confirms cross-country patterns of central tendency to be robust to choice of model. But the same is not true for dispersion. This is particularly the case if one looks at a group of countries at widely diverging levels of development. Results for less developed countries are much less robust. 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 a conclusion over the broad direction of association between central tendency and dispersion appeared sensitive to choice of model when we looked at the group of all countries who participated in TIMSS in 1995, irrespective of their level of development. We believe that survey reports should include analysis of the sensitivity of basic results to model choice of the type we have made here, moving this area of social science measurement more in line with others where concern over robustness of results to choice of model or measurement method is more apparent (e.g. Mroz 1987 on labour supply and Coulter et al 1992 on income inequality and poverty).

Appendix

i) country definitions

UK. Our TIMSS data for the UK refer only to England and Scotland; the data for England are drawn from TIMSS 1999 and are combined (with appropriate weights to account for differences in population size) with data (three-parameter scores) for Scotland drawn from TIMSS 1995 (Scotland did not participate in TIMSS 1999). Likewise, PIRLS data for the UK refer to England and Scotland only. For PISA, the UK is represented by England, Scotland and Northern Ireland. IALS covers all parts of the UK.

Belgium. We combine TIMSS 1999 data for Flemish speaking areas with 1995 data (three parameter scores) for French speaking areas. IALS data refers to Flanders only.

Canada. PIRLS coverage is restricted to the provinces of Ontario and Quebec.

Norway. IALS results are restricted to speakers of Bokmal Norwegian, the language of the large majority of Norwegians.

ii) Standard deviation of the average z-score

The variation in average z-scores in Tables 2 and 3 may be used to assess the degree of agreement between tests. If the z-scores for all countries were the same in each test then the standard deviation of the average z-scores would be equal to 1.0. If the tests were completely unrelated its expected value would be 0.353 (= $[1/8]^{½}$). In reality the value lies half-way in between these two extremes for both the median and P95-P5: 0.78 and 0.74 respectively. How can these values be interpreted? Assume that a country's z-score for each test is the sum of a component specific to that country and common to all tests and another that is independent for each country and each test. Assume also that these components have the same variance. In other words, there is a country component that is the same for all tests and one that differs from test to test, and they are of equal importance (in variance terms). In this situation, the standard deviation of the average z-scores would be equal to 0.75, were each test weighted equally in the calculation. (The values noted above of 0.78 and 0.74 are based on weighting the surveys equally and not the tests but if we switch to the latter basis the values hardly change.) Hence the observed values of the standard deviations are consistent with the relative importance of the two components stated above, given the other assumptions made. These assumptions are unlikely to be true and the calculation should be interpreted as a summary statistic of the degree of similarity of the test results and not as the estimate of a structural parameter. The eight tests in Tables 2 and 3 come from only three surveys, a fact ignored by the assumption that the second component of the country z-score is independent across tests.

References

Adams R J (2002), 'Scaling PISA Cognitive Data' in R Adams and M Wu (eds.) *PISA 2000 Technical Report*, OECD, Paris.

Adams R J (2003), 'Response to 'Cautions on OECD's recent educational survey (PISA)', *Oxford Review of Education*, vol. 29, no 3.

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

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

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

Beaton A (2000), 'The importance of Item Response Theory (IRT) for large scale assessments' in Carey S (ed.), *Measuring Adult Literacy. The International Adult Literacy Survey (IALS) in the European Context*, Office for National Statistics, London.

Blum A., H. Goldstein and F. Guerin-Pace (2001), 'An analysis of international comparisons of adult literacy', *Assessment in Education*, vol. 8, no. 2, pp.225-246.

Brown G and J Micklewright (2004), 'Using International Surveys of Achievement and Literacy: A View from the Outside', Working Paper 2, UNESCO Institute for Statistics, Montreal.

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

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

Coulter F, Cowell, F and Jenkins S (1992) 'Equivalence Scale Relativities and the Extent of Inequality and Poverty' *The Economic Journal*, vol 102, no 414, pp.1067-82.

Denny K. (2003), 'New Methods for Comparing Literacy across Populations: Insights from the Measurement of Poverty', *Journal of the Royal Statistical Society: Series A*, vol. 165, issue 3, pp. 481-493.

Esping-Andersen, G. (2004), 'Unequal Opportunities and the Mechanisms of Social Inheritance', in M. Corak (editor), *Generational Income Mobility in North America and Europe*, Cambridge: Cambridge University Press.

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

Goldstein H. (2004), 'International comparisons of student attainment: some issues arising from the PISA debate' forthcoming *Assessment in Education*.

Gonzalez E. (2003), 'Scaling the PIRLS Reading Assessment Data' in M Martin, I Mullis, A Kennedy, (eds.) *PIRLS 2001 Technical Report,* Chestnut Hill, MA: Boston College.

Micklewright J. and S.V. Schnepf (2004), 'Educational achievement in English-speaking countries: do different surveys tell the same story?' Applications and Policy Working Paper A04/10, Southampton Statistical Sciences Research Institute, University of Southampton.

Micklewright J. and S.V. Schnepf (2005), 'How good is education in Central and Eastern Europe?', forthcoming working paper, UNESCO Institute for Statistics, Montreal.

Mroz, T (1987), 'The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions', *Econometrica*, 55(4):765-799

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

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

Prais, S.J. (1997). 'Whole-class teaching, school-readiness and pupils' mathematical attainments', *Oxford Review of Education*, vol. 23, no. 3, pp. 275–90.

Prais, S J (2003), 'Cautions on OECD's recent educational survey (PISA)', *Oxford Review of Education*, Vol. 29, pp. 139-163.

OECD and Statistics Canada (2000), *Literacy in the Information Age – Final Report of the International Adult Literacy Survey*, OECD, Paris.

OECD (2001), *Knowledge and Skills for Life – First results from PISA 2000*, OECD, Paris.

OECD and UNESCO Institute for Statistics (2003*) Literacy Skills for the World of Tomorrow - Further results from PISA 2000*, OECD, Paris.

Social Exclusion Unit (2001), *Preventing Social Exclusion*, Social Exclusion Unit.

UNDP (2000), *Human Development Report*, UNDP, New York

Yamamoto K (1998), 'Scaling and Scale Linking' in T Scott Murray, I Kirsch, and L B Jenkins (eds.) *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey*, National Center for Education Statistics, Washington DC

Yamamoto K and E Kulick (2000), 'Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales' in M Martin, K Gregory, and S Stemler (eds) *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

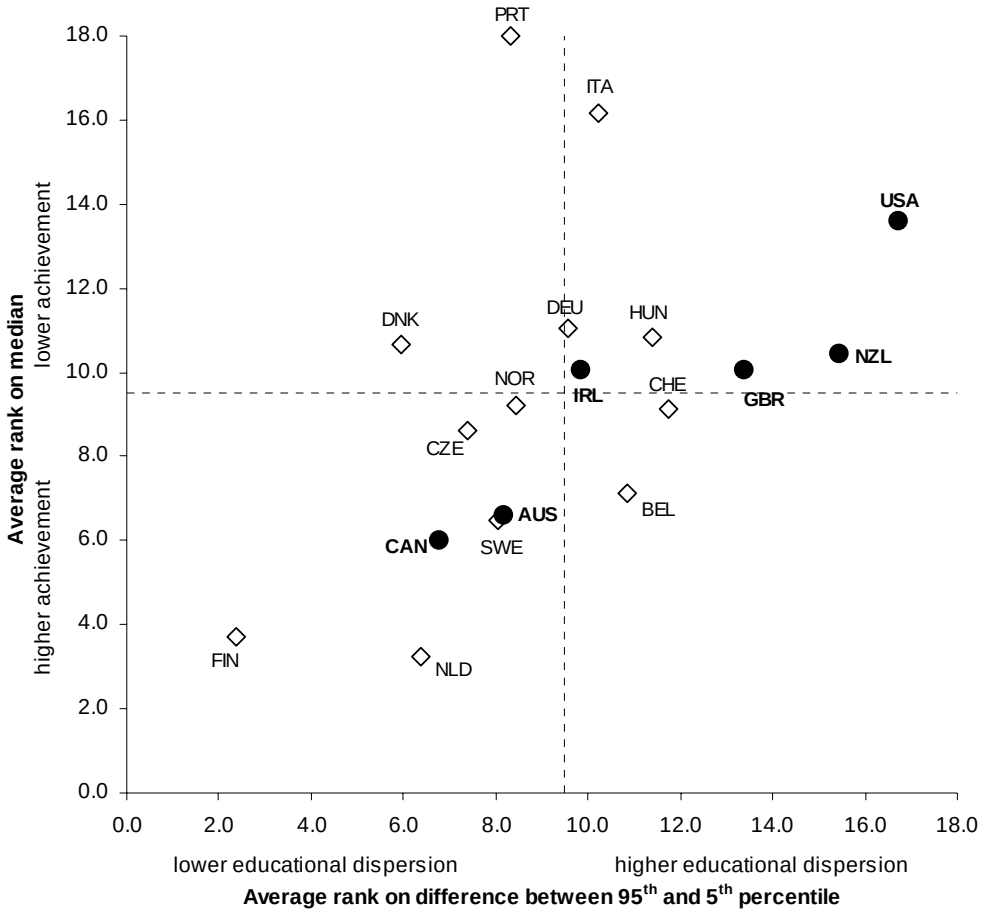
Wößmann L. (2003), 'Schooling Resources, Educational Institutions and Student Performance: the International Evidence', *Oxford Bulletin of Economics and Statistics*, vol. 65(2), pp. 117-170.

Table 1: Cross-national survey data on achievement used in this paper

Survey	Round	Age group	Subjects covered	Sample size per country (av.)
Trends in International Maths and Science Study (TIMSS)	1995 1999	13 - 14 (grade 8)	maths and science	3,800
Programme of International Student Assessment (PISA)	2000/2002	15	reading, maths and science	5,700
International Adult Literacy Survey (IALS)	1994-98	16-24	document, prose and quantitative literacy	700
Progress in International Reading Literacy Study (PIRLS)	2001	9 -10 (grade 4)	reading	4,300

Notes: TIMSS and PIRLS are organised by the International Study Center, Boston College, USA. PISA is organised by OECD. IALS was organised by OECD and Statistics Canada. The first PISA survey of 2000 was repeated in further countries in 2002 (PISA+). PISA+ data are included in our analysis.

Figure 1: Average rank on central tendency (median) and dispersion (P95-P5) for 18 countries in 8 tests (PISA, TIMSS, IALS)



Note: the graph shows the average rank on the median and the difference between 95th and 5th percentiles in 3 surveys (TIMSS, PISA, IALS) with 8 tests for 18 countries. The higher the median and the lower the dispersion the smaller in number the rank. Gridlines show the average for all countries (9.5). The surveys are equally weighted in the averaging.

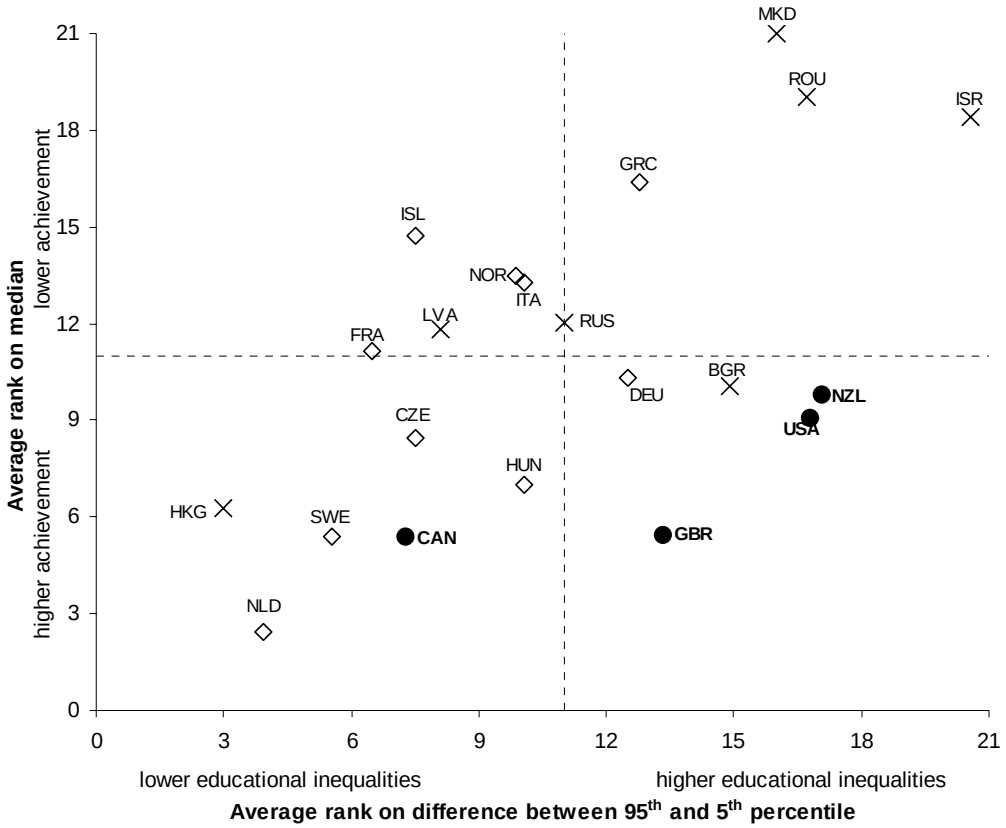
Table 2: Average ranks and z-scores for the median for 18 countries covered in PISA, TIMSS and IALS

	Average rank	Average z-score	PISA			TIMSS		IALS		
			read	maths	science	Maths	Science	doc	quant	prose
Netherlands	3.2	1.09	1.11	1.97	1.12	1.26	1.26	0.77	0.50	0.72
Finland	3.7	1.11	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	0.69	0.32	-0.03	0.33	0.13	0.17	1.34	1.43	1.84
Australia	6.6	0.48	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 Rep.	8.6	0.13	-0.84	-0.56	0.12	0.15	0.75	0.36	1.18	-0.16
Switzerland	9.1	0.16	-0.62	0.67	-0.62	1.12	-0.26	0.52	0.55	-0.10
Norway	9.2	0.18	-0.06	-0.40	-0.22	-0.50	-0.22	1.14	0.76	0.99
UK	10.1	0.05	0.51	0.60	1.12	-0.73	0.63	-0.37	-0.93	-0.46
Ireland	10.1	-0.06	0.78	-0.23	0.24	0.45	0.04	-0.85	-0.59	-0.29
New Zealand	10.4	0.02	0.99	0.95	1.06	-0.85	-0.30	-0.60	-0.85	-0.23
Denmark	10.7	-0.25	-0.53	0.12	-1.12	-0.44	-1.95	0.90	0.97	0.05
Hungary	10.8	-0.40	-1.39	-0.93	-0.50	0.90	1.51	-1.16	-0.14	-1.45
Germany	11.1	-0.24	-1.00	-0.61	-0.84	-0.27	0.22	0.08	0.58	-0.11
USA	13.6	-0.72	-0.25	-0.62	-0.33	-0.39	-0.10	-1.20	-1.77	-1.12
Italy	16.2	-1.18	-1.09	-1.83	-1.30	-1.29	-1.16	-1.13	-1.07	-0.61
Portugal	18.0	-2.10	-1.85	-1.96	-2.23	-2.60	-2.19	-2.12	-1.95	-1.89

Table 3: Average ranks and z-scores for P95-P5 for 18 countries covered by PISA, TIMSS and IALS

	Average rank	Average z-score	PISA			TIMSS		IALS		
			read	maths	science	Maths	science	doc	quant	prose
Finland	2.4	-1.35	-1.45	-1.76	-1.68	-1.70	-0.87	-0.99	-1.47	-0.93
Denmark	5.9	-0.52	-0.19	-0.96	0.93	-0.19	0.04	-1.29	-1.12	-1.78
Netherlands	6.4	-0.75	-1.38	-0.94	-0.02	-0.51	-1.01	-0.92	-0.47	-0.79
Canada	6.8	-0.48	-0.58	-1.17	-1.27	-0.62	-0.86	0.82	-0.09	0.17
Czech Rep.	7.4	-0.28	-0.23	0.62	-0.42	0.23	-0.53	-0.22	-0.81	-1.04
Sweden	8.1	-0.60	-0.86	0.15	-0.68	-1.08	-1.05	-0.42	-0.38	-0.01
Australia	8.2	0.03	0.34	-0.30	-0.44	0.28	0.41	-0.39	-0.05	0.04
Portugal	8.3	-0.72	-0.18	-0.27	-1.41	-1.65	-1.29	-0.23	-0.47	0.47
Norway	8.4	-0.47	0.72	-0.12	-0.22	-0.40	-1.37	-0.47	-0.52	-0.95
Germany	9.6	0.33	1.88	1.38	0.95	-0.06	1.13	-1.10	-1.14	-0.63
Ireland	9.8	0.13	-0.65	-1.38	-0.83	0.75	0.88	0.35	0.83	0.40
Italy	10.2	0.14	-1.19	-0.28	0.13	1.22	0.56	-0.25	0.06	0.15
Belgium	10.8	0.55	1.21	1.88	2.41	0.21	0.08	-0.85	-0.09	-0.05
Hungary	11.4	0.36	-0.79	0.68	0.78	1.10	0.02	0.61	0.60	-0.36
Switzerland	11.7	-0.04	0.50	0.96	0.42	-1.15	-0.52	0.14	-0.08	0.20
UK	13.4	0.80	0.28	-0.15	0.27	0.70	1.13	1.32	1.63	1.08
New Zealand	15.4	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

Figure 2: Average rank on median and P95-P5 in 6 tests for 21 countries (PISA, PIRLS, TIMSS)



Note: Surveys equally weighted. Gridlines show average rank of all countries. Non-OECD countries are indicated by a cross.

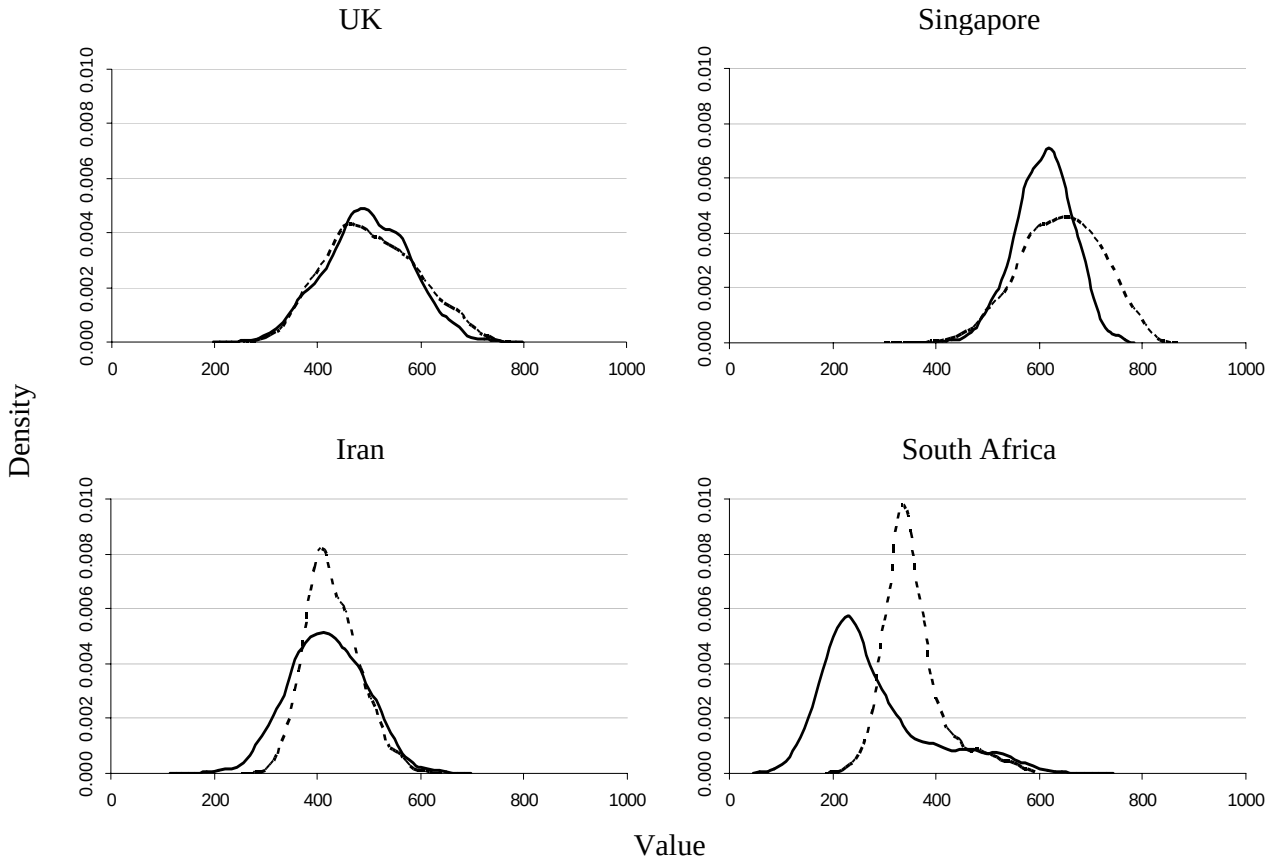
Table 4: Correlation matrix of median for 18 countries covered by PISA, TIMSS and IALS

			PISA			TIMSS		IALS		
			Reading	Maths	Science	Maths	Science	Prose	Doc.	Quant.
Q50	**PISA**	Reading	1							
		Maths	0.82	1						
		Science	0.90	0.80	1					
	TIMSS	Maths	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		
		Doc.	0.50	0.61	0.46	0.54	0.25	0.91	1	
		Quant.	0.21	0.40	0.24	0.59	0.28	0.74	0.89	1
Q95-Q5	PISA	Reading	1							
		Maths	0.73	1						
		Science	0.57	0.73	1					
	TIMSS	Maths	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		
		Doc.	0.25	0.17	0.00	0.56	0.55	0.87	1	
		Quant.	0.28	0.23	0.23	0.70	0.67	0.88	0.91	1

Table 5: Correlation matrix of median and Q95-Q5 for 21 countries covered by PISA, TIMSS and PIRLS

			PISA			TIMSS		PIRLS
			Reading	Maths	Science	Maths	Science	Reading
Q50	PISA	Reading	1					
		Maths	0.94	1				
		Science	0.96	0.96	1			
	TIMSS	Maths	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
Q95-Q5	PISA	Reading	1					
		Maths	0.56	1				
		Science	0.57	0.63	1			
	TIMSS	Maths	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

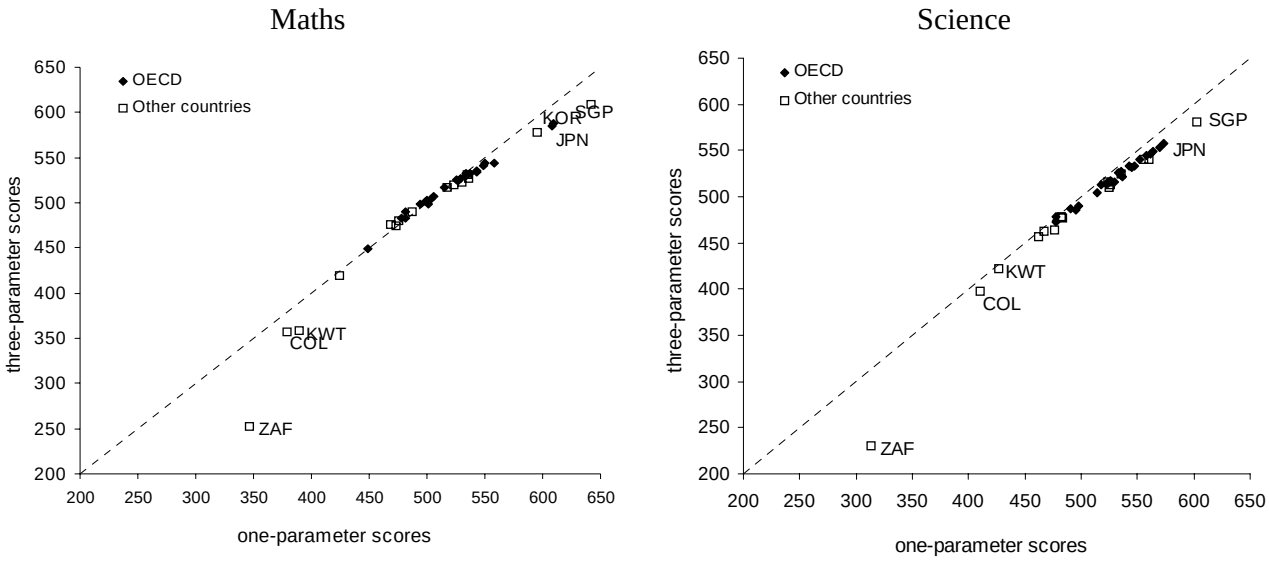
Figure 3: Distribution of 8th graders' achievement in maths in TIMSS 1995



-- one-parameter scores
— three-parameter scores

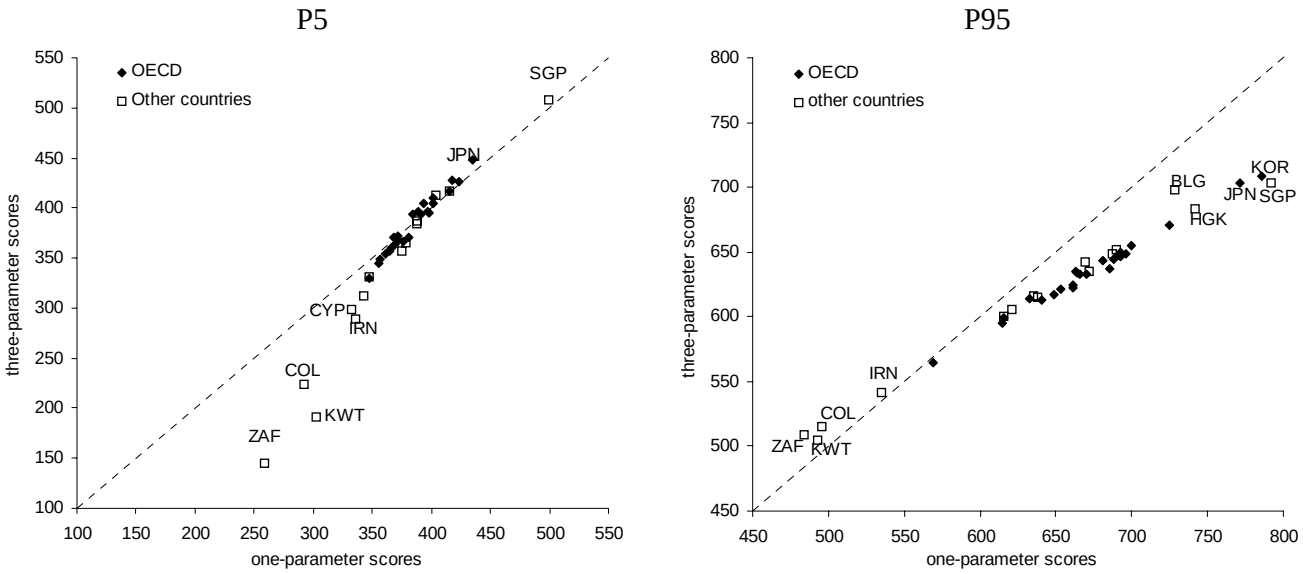
Note: the distributions depicted are of the averages of the five plausible values for each individual.

Figure 4: Comparison of medians of one-parameter and three-parameter values



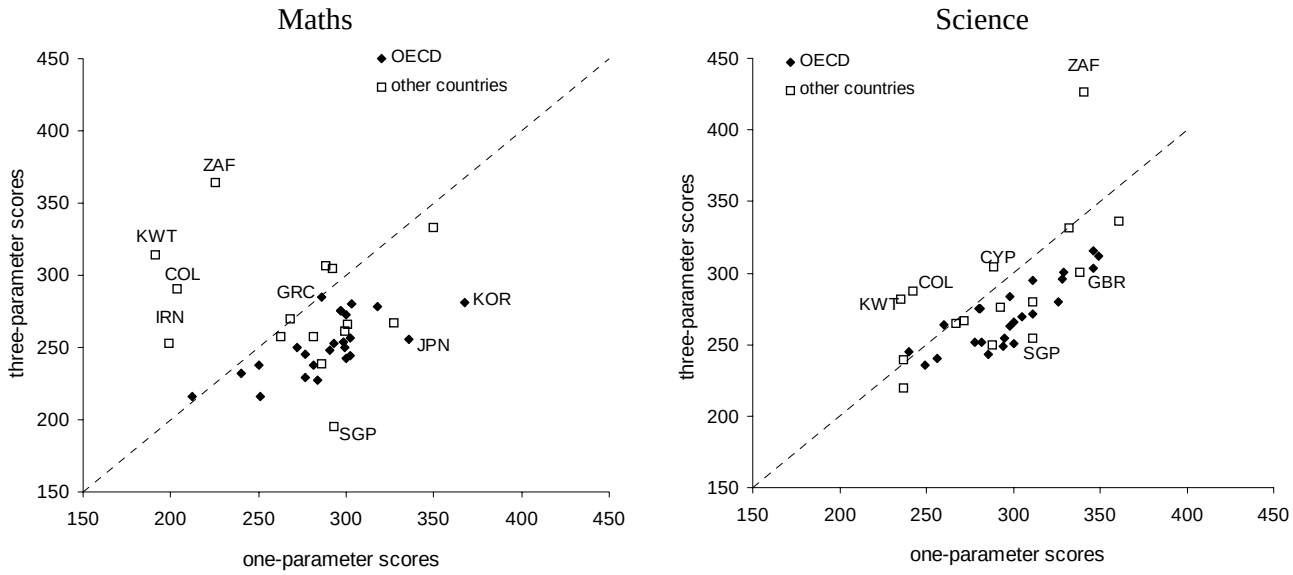
Note: the correlations of one- and three-parameter medians are 0.98 for maths (1.00 for OECD countries) and 0.97 for science (0.99 for OECD countries).

Figure 5: Comparison of P5 and P95 of one-parameter and three-parameter values in maths



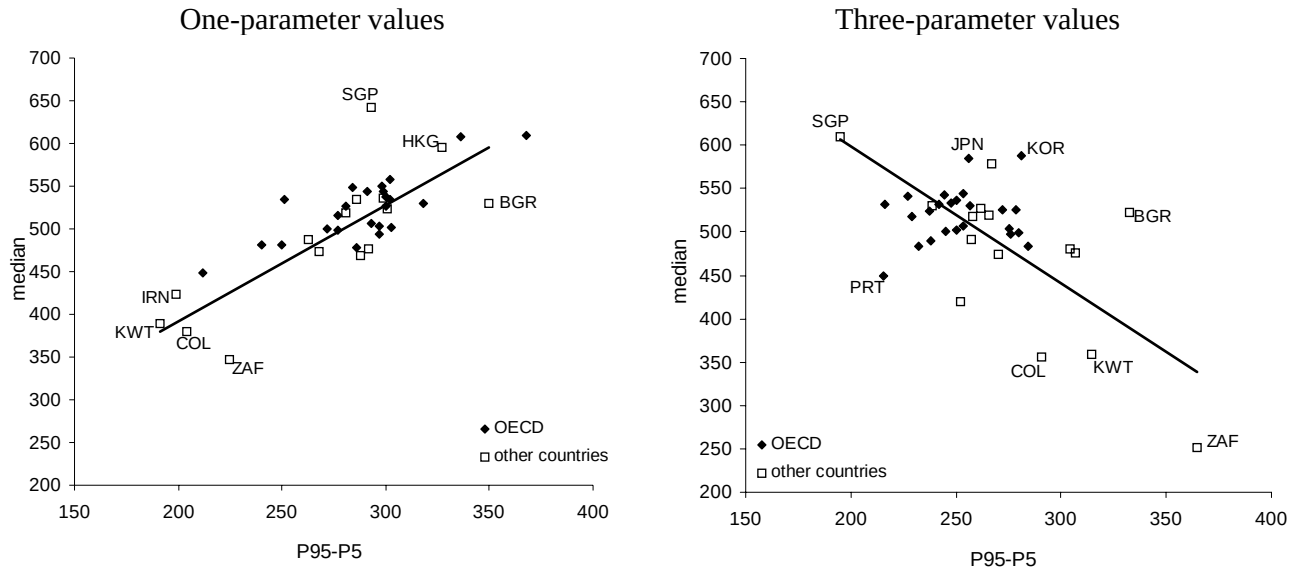
Note: the correlations of one- and three-parameter values are 0.97 for P5 (0.98 for OECD countries) and 0.99 for P95 (1.00 for OECD countries).

Figure 6: Comparison of P95-P5 of one-parameter and three parameter values



Note: the correlations of one- and three-parameter values of P95-P5 are 0.03 for maths (0.70 for OECD countries) and 0.67 for science (0.85 for OECD countries).

Figure 7: Association of P95-P5 and P50 with different item response models



Note: the correlations of 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).

Table 6: Correlation of one-parameter and three-parameter values of median and P95-P5 in TIMSS 1995 with PISA values

	All 29 countries		23 OECD countries	
	Median	P95-P5	Median	P95-P5
TIMSS 3 parameter	0.55	0.40	0.70	0.17
TIMSS 1 parameter	0.54	0.18	0.69	0.01