INEQUALITY OF LEARNING IN INDUSTRIALISED COUNTRIES

JOHN MICKLEWRIGHT, SYLKE V. SCHNEPF

ABSTRACT

Within-country differences in educational outcomes are compared for a large group of industrialised countries. We investigate where inequality is greatest, the association between inequality in learning and average levels of learning, the interpretation of measured levels of inequality, and differences in inequality at the top and bottom of the national distributions. Our analysis is based on test score data for 21 countries present in the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS). The use of three different surveys avoids reliance on a single source.

Southampton Statistical Sciences Research Institute
Applications & Policy Working Paper A06/07

Inequality of learning in industrialised countries

John Micklewright* and Sylke V. Schnepf**

* Southampton Statistical Sciences Research Institute (S3RI) and School of Social Sciences, University of Southampton.

** S3RI, University of Southampton

December 2006

Abstract

Within-country differences in educational outcomes are compared for a large group of industrialised countries. We investigate where inequality is greatest, the association between inequality in learning and average levels of learning, the interpretation of measured levels of inequality, and differences in inequality at the top and bottom of the national distributions. Our analysis is based on test score data for 21 countries present in the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS). The use of three different surveys avoids reliance on a single source.

To be published in S. P. Jenkins and J. Micklewright (eds.) *Inequality and Poverty Re-examined*, Oxford University Press, 2007.

Key words: inequality, learning, education, TIMSS, PISA, PIRLS
JEL codes: D39, I21, I39

Acknowledgments

This paper stems from a research project undertaken with Giorgina Brown, of ISTAT, Rome, and Robert Waldmann, of University of Rome ‘Tor Vergata’, and we draw in part on unpublished work with them. The research was partly funded by UNESCO Institute for Statistics, Montreal (the views expressed are our own and should not be associated with UNESCO). We are grateful for comments to Stephen Jenkins and to seminar participants at the Institute for Fiscal Studies, London, and at Cornell and Princeton.

1. Introduction

The importance of human capital in determining incomes leads quickly to interest in the extent of educational inequalities. Differences in education have a major impact on the distribution of earnings and on the number and characteristics of the poor. Inequalities in education also help produce disparities in well-being in dimensions other than income. These include both obvious dimensions such as improved health and higher occupational status, and less obvious ones such as ability to perceive and take advantage of a range of opportunities: 'the educational level of the retired, for example, is relevant to their capacity to participate in society and to take part in the democratic process' (Atkinson *et al*. 2002: 128).

Our aim in this chapter is to compare within-country differences in educational outcomes across a large group of industrialised countries. Where are these differences greatest? We are immediately faced with the issue of how to measure levels of education. One option would be to focus on data on 'attainment', that is on levels of education that have been completed (or at least entered): primary, secondary, tertiary etc. There are significant literatures within both economics and sociology that use this form of information to compare educational inequalities across countries. For example, Thomas, Wang and Fan (2001) compare the distribution of the population across seven levels of attainment for 85 developing and industrial countries for the period 1960–90, attributing a given number of years of schooling to each level. Other authors compare social class differences in attainment across countries, for example Müller (1996) and Shavit and Blossfeld (1993).

We take a different route, focusing on survey data that record what people actually know, as measured by performance in tests. These are a form of 'achievement' data. Recent years have seen several international surveys of learning achievement of children and 'functional' literacy of adults (the ability to function in modern society). Samples of individuals are administered standardised tests with the aim of comparing countries' levels of achievement or literacy and the factors that influence them. These surveys, with their purpose-built design for cross-national comparison, offer the hope of cutting through the problems of comparing national educational systems that are presented by attainment data.

But which achievement survey to use? There is the International Adult Literacy Survey (IALS), the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS). Each survey aims to assess something different or to assess knowledge in a different way. They each refer to particular age groups or school grades. And they each have been the subject of criticism on one ground or another. Our main contribution is to compare results across the surveys. This contrasts with the typical analysis of the data, whether of inequalities or of any other aspect of achievement, that is restricted to a single source. We use TIMSS, PISA and PIRLS.

These three surveys all refer to children of compulsory school age. We are therefore comparing differences in educational outcomes across countries that emerge *before* the decisions at the end of compulsory schooling and in the ensuing years that generate most of the variation in attainment in industrialised countries. These decisions, both by individuals and their families and by schools, colleges and universities, are strongly influenced by learning that has taken place during the compulsory school period. This learning also has direct effects on wellbeing in adult life.

We next describe the data and the tools we use to compare within-country differences in achievement scores. These tools are simple and allow for the fact we have multiple sources. We also take into account the nature of the achievement data, which are very different to income data. Our main results follow in which we focus on a group of 21 countries present in all three surveys. We investigate where inequality is greatest, the association between inequality in learning and average levels of learning, the interpretation of measured levels of inequality, and differences in inequality at the top and bottom of the national distributions. In the concluding section we discuss future directions for the analysis of educational inequalities with achievement data.

2. Data and tools

The international achievement surveys

TIMSS, PISA and PIRLS have similar sample designs. They all involve the selection of a sample of schools and then a single class (TIMSS and PIRLS) or a random sample (PISA) of pupils within each school. Typical sample size in any country is about 150 schools and about 30 pupils per school. TIMSS is perhaps the best well known. We use data on children in grade 8 (usually aged 13–14) from the 1995 and 1999 rounds of the survey (taking the data from the later year if a country participated in both rounds).[1] The PISA data relate to an age group – 15 year olds – rather than a grade. We use data from the 2000 round. PIRLS focused on children

[1] About one third of the questions in 1999 were the same as in 1995. The others were intended to give results that were comparable. The precise selection of data is described in Brown *et al*. (2007).

in grade 4 (usually aged 9–10) and we use the data from the first round of the survey which was held in 2001.

The surveys differ in a number of ways other than the differences in the target population.[2] Notably, they vary widely in the type of achievement that they try to assess. PISA assesses ability in reading, science and maths, attempting to determine to what extent 'education systems in participating countries are preparing their students to become lifelong learners and to play constructive roles as citizens in society' (OECD 2001). (Note that 'education systems' should be interpreted as the combination of schools and families and not just schools.) The aim is to measure broad skills, and to see how students would be able to use what they have learned in real-life situations. While covering a similar age group to PISA and two of the same subjects – maths and science – TIMSS focuses more on measuring mastery of internationally agreed curricula. This may seem a narrow approach. But at least the concept of a curriculum agreed by educationalists is one that a lay person can begin to understand, even though the content of that curriculum is subject to debate. By contrast, at first sight the 'life-skills' approach of PISA may seem more nebulous. It may also be easier to carry out measurement of achievement against a standard in a culture-free way in TIMSS. PIRLS measures only achievement in reading. The survey organisers argue that their approach is similar to that in PISA, both being based on 'an expanded notion of literacy' (Campbell *et al*. 2001: 85).

These three sources provide information on achievement in a total of six tests. We restrict attention to the 21 countries present in all three surveys. This group comprises 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). Our findings therefore relate to a rather arbitrary group of countries, but one that is not dissimilar to, for example, those present in the Luxembourg Income Study that is widely used to analyse inequality in incomes.

The answers that a respondent gives to the questions in the surveys are summarised by the organisers into a single score for the subject concerned – maths, science, reading – using an 'item response' model. The purpose of the modelling is to estimate the unobserved distribution of proficiency in a subject from the observed answers to the test questions. While the raw scores to the test questions have a theoretical maximum, the unobserved proficiency distribution is unbounded and one of the purposes of the modelling is to allow for the implied

[2] See Brown *et al*. (2007) for more discussion.

censoring of high achievement in the raw scores. (See, for example, Beaton 2000.) The steps in the process used in each survey are similar but the precise model that is employed differs from survey to survey. Scores for each country are scaled by the survey organisers to have a mean among all persons in all participating countries (which is always a wider group than the 21 countries present in all three surveys that we consider here) of 500 points and a standard deviation of 100 points.[3]

Measuring inequality in learning

The achievement test data are recorded on a continuous scale. This suggests that in measuring inequality of learning we could select from the full range of tools that have been developed to measure inequality in incomes, and the differences in this inequality across countries. The tools of income inequality analysis have, after all, been applied in international comparisons of other non-income dimensions of well-being, including height as a proxy for health (for example Pradhan, Sahn, and Younger 2003). And they have also started to enter the analysis of the international education surveys of the type used here. Denny (2002) uses methods developed for the measurement of poverty to analyse levels of low functional literacy in IALS, including, for example, Foster-Greer-Thorbecke indices.

We hesitate over use of more sophisticated measures, for three reasons. First, there is the practical problem of multiple sources of data. We draw on three surveys covering six different test distributions. This amount of information would considerably complicate any dominance analysis, for example, in which one tried to order (at least partially) the country distributions in a way that would be independent of the choice of a particular inequality index. One of our aims when using multiple sources is to condense the information they contain and as a result we are drawn towards summary measures of learning inequality for each country that are easier to compare across tests and surveys.

Second, the nature of the achievement test data calls for caution in the use of the income inequality measurement toolbox. The test scores are *derived* data providing estimates of proficiency in different subjects. It is doubtful whether the measurement of the scores is on a ratio scale. Their nature is therefore quite different from that of data on income or height. The fact that scores are scaled by the survey organisers to have the same mean and standard

[3] Survey organisers use the item response models to produce what in fact are five ‘plausible values’ of proficiency for each individual rather than a single figure. We follow the organisers’ practice of calculating all summary statistics of the score distributions with each plausible value and then averaging the five resulting estimates.

deviation does not make them inherently comparable across tests. The choice of item response model influences the shape of the estimated proficiency distributions and can do so in ways that change the cross-country picture. Together with Giorgina Brown and Robert Waldmann, we have shown that rankings of countries by within-country differences in TIMSS changed quite sharply in some cases when the survey organisers applied retrospectively the model used in the 1999 survey round to the 1995 data, although the changes are much less when low income countries that we do not use in this chapter are excluded (Brown *et al*. 2007).[4] We are therefore reluctant to compare directly the *levels* of inequality indices of learning across the different tests since these are in part a function of the chosen item response model.[5]

Third, given the lack of much previous work focusing directly on learning inequality as measured by the achievement surveys, we want to explore the shape of the test score distributions in a little more depth than is made possible by the use of a single index. We therefore focus on quite crude measures: differences in quantiles of the test score distributions. We consider the 95th percentile minus the 5th percentile, P95–P5, the 95th minus the 50th, P95–P50, and the 50th minus the 5th, P50–P5. The use of the latter two measures can reveal whether differences in inequality across countries are more obvious in the top half or in the bottom half of the range of scores. (Any answers, of course, are conditional on the particular item response model used to produce the distributions in question.) We allow for sampling variation when comparing these measures across countries (taking into account the complex survey designs).[6] These measures of absolute differences in scores contrast with indices of inequality that are most commonly used in the analysis of incomes, which relate to relative differences, for example quantile ratios or the Gini coefficient. However, we have no particular reason to focus on relative as opposed to absolute differences, especially since we doubt whether our data measure achievement on a ratio scale. And by not presenting ratios of scores, we remove any temptation to try to compare levels of inequality in the achievement data with those shown by quantile ratios for earnings or income distributions, a comparison that we feel would not be valid.

[4] We use the re-modelled 1995 data for countries that did not take part in the 1999 round. These data are in principle consistent with those for 1999.

[5] We are certainly not the first to be cautious with data measuring achievement or ability. Atkinson (1975: 89) notes the comment of Mayer (1960) that 'there is at present really no such thing as *the* distribution of ability: the distribution depends on the measuring rod used and cannot be defined independently of it'. Atkinson warns 'the fact that most IQ tests lead to a distribution of scores which follows the normal distribution does not necessarily tell us anything about the distribution of abilities: it may simply reflect the way in which the tests have been constructed'.

[6] The derivation of standard errors of the quantile differences is described in Brown and Micklewright (2004) and uses the survey organisers' estimates of the standard errors of the quantiles. These estimates are not provided for the 10th and 90th percentiles, helping determine our choice of the 5th and 95th percentiles.

3. Results

Where is inequality highest?

Table 1 shows the values of P95–P5 in each of the six tests. These values have been transformed into z-scores. That is, for each test we adjust a country's value of P95-P5, expressed in points of achievement score, by subtracting the mean value of P95–P5 for the 21 countries that we consider and by dividing by the standard deviation of the 21 values. The z-scores therefore show how far each country is above (positive values) or below (negative values) the mean value of P95–P5 for the test in question, expressed in standard deviations. This transformation represents our best effort to reduce problems of comparability across tests that are presented by the nature of the achievement score data. The z-scores for each test are shown in columns 3-8.

The first two columns show for each country its average z-score (column 2) and its average rank (column 1) in the orderings implied by the values in columns 3-8. The surveys, and not the tests, are weighted equally in calculating these averages. (This implies, for example, that the scores and ranks for each of the three PISA tests receive one third of the weight given to the scores and ranks for the single PIRLS test.) These averages have some merit as quick summary statistics. If the different tests were to produce sharply differing orderings of the countries, the averaging would produce figures with little variation. A negative z-score or low rank in one 'league table' would likely be balanced by a positive z-score or high rank in another. The more the average z-scores and ranks vary the more the different tests must be in agreement. Having a low or high average rank can only result from ranking consistently well or consistently badly in individual tests. The countries are ordered in the table on the basis of the average z-scores (which corresponds closely to the ordering on average ranks). The shading in columns 3 to 8 indicates the third of the distribution of values for that test into which a country falls: dark shading for the seven countries with the largest values of P95–P5 for the test concerned, light shading for the seven countries with intermediate values, white for the seven countries with the smallest values.

The table shows that there is a reasonable degree of agreement between the six tests, taken as a whole: the average z-scores and average ranks *do* vary considerably. Leaving aside the maximum and minimum values, the average z-scores range from −1.1 to +1.0 and the average rank from 3.8 to 17.1. Hong Kong, the Netherlands, Sweden, France, Canada and

Iceland are the countries where the within-country differences tend to be smallest and Bulgaria, the USA, Macedonia, New Zealand, Romania and Israel the countries where the biggest differences are found. Israel is a clear outlier, with the largest differences in most tests and often by a large margin. The group of the most unequal countries therefore includes both the poorest member of the 21 country pool, Macedonia, and one of the richest, the USA. It is notable that the group contains three Eastern European countries. The future path of educational inequalities in these countries is an obvious subject to track in further rounds of the international achievement surveys.[7]

The degree of agreement between the different tests, and in particular the different surveys, is encouraging. However, there is also some disagreement between the figures and this means that caution is required when looking at results from just one source. Israel aside, no country is in the same third of the distribution of countries for all six tests, with the same shading throughout columns 3 to 8. Nine of the 21 countries have three different shadings. For example, inequality in learning in the UK is well below the average for maths in PISA but well above the average for reading in PIRLS, with values in the middle third of countries for the other four tests. Russia has below average values in both the PISA and PIRLS reading tests but the opposite in both the PISA and TIMSS maths and science assessments. This underlines the importance of looking at the whole set of tests before coming to any conclusions about a country's position.

When comparing results across the tests, one possibility is that differences may reflect the different age groups covered by the surveys. In particular, PIRLS surveys children of primary school age but PISA and TIMSS survey children of secondary school age. Factors affecting achievement may vary with age. Primary school systems are typically comprehensive, with schools all containing a mix of children of different abilities, but secondary schooling in some countries divides children into separate academic and technical schools. This may have the effect of increasing inequalities in achievement (see, for example, Wößmann 2003). For example, it is notable that Germany, a country that does separate children into quite different school types straight after primary schooling, has a below average value of P95–P5 for reading for the 9–10 year olds in PIRLS but the highest value of all the 21 countries for reading for the 15 year olds in PISA.

However, in general the comparison of a country's position between any two tests needs to consider other factors as well. In principle, the difference between Germany's

[7] Inequalities in education in Central and Eastern Europe and the impact of economic and social change following the end of the communist period are discussed in Micklewright (1999) and UNICEF (2001).

position in PISA and PIRLS might conceivably reflect something to do with the different nature of the two assessments, or the item response modelling, or the organisation of the surveys in Germany. (And it should also be noted that Germany does not stand out in Table 1 in TIMSS or, especially, in the other two PISA subjects.) Why is it, to take a very different example, Macedonia is the most unequal country in reading in PIRLS by some way, but has below average inequality in reading in PISA, and is the least unequal country of all in PISA science? The fact that there is much more agreement for Macedonia between PIRLS and TIMSS than between PIRLS and PISA suggests that the differences between the latter two surveys reflect more than age.

The broad agreement between the different tests, as reflected in the extent of variation in the average ranks and z-scores, implies that the observed patterns in the data are unlikely to reflect merely the noise from sampling variation. We can test this formally: taking all pairwise comparisons of P95–P5 between countries, about a half are significant at the 5 percent level in the PISA and TIMSS tests and about three-quarters in PIRLS.[8] Naturally, differences between countries that are close in the ranking on any test are typically insignificant. But in PIRLS, for example, the values of P95–P5 for the eight countries with the smallest values are all significantly different from each of the values for the eight countries at the opposite end of the ordering.

Inequality in learning versus average learning

It is of obvious interest to investigate the relationship between inequality of learning and average levels of achievement. Are the countries where inequalities in education are smallest also those where average levels are greatest? Or does there appear to be a trade-off, so that a focus on reducing within-country differences has the effect of depressing the average? In their large cross-country study of attainment, measured by years of schooling, Thomas, Wang and Fan (2001) find that the former pattern clearly holds: inequality falls as average years of education rises. This finding relates to inequality measured using the scale-invariant Gini coefficient. When the authors switch to using the standard deviation (which is not scale invariant), an inverted U-shape Kuznets curve for education is found, with inequality first rising with average years of schooling and then falling. The turning point comes at about 6 to 7 years of schooling, a level exceeded by industrialised countries. Hence for these countries,

[8] We do not apply the Bonferroni correction for multiple comparisons.

their results show inequality in attainment is inversely related to average attainment, irrespective of whether a scale-invariant measure of inequality is used or not.

In contrast to Thomas *et al*., who had repeated observations on the same countries over time, we cannot investigate this issue adequately with our achievement data. All we can do is to use a single observation – based on the three surveys we use – for our cross-section of countries to show the association of inequality in learning with average levels. Figure 1 plots the average z-scores of P95–P5 given in column 2 of Table 1 against the average z-scores across the same six tests for P50, the median. (The z-score transformations for the medians are performed in an analogous fashion to those for P95-P5; the z-scores measure how far a country is above or below the mean value of P50 for the 21 countries, measured in units of standard deviations of P50.) There is a reasonably clear pattern: in broad terms, within-country differences are highest (positive z-scores) where average achievement is lowest (negative z-scores).

This accords with the pattern found for industrialised countries by Thomas *et al*. with their very different data.[9] (The use of scale-invariant measures of relative differences for the achievement data, such as the ratio of P95 to P5, would emphasise this pattern even more.) This said, our results for particular countries do not always reflect the position found in the attainment data by Thomas *et al*. The USA appears as the most *equal* country in their study but one of the most unequal according to our achievement data. We also note that our work comparing results in TIMSS for different item response models (Brown *et al*. 2007) shows that the association of score dispersion and central tendency is not always robust to the choice of model, especially when low income countries (excluded here) are included in the analysis. And with both types of data, attainment and achievement, the issue arises of whether the association between inequality and the average is simply a result of the nature of the data, or, to be more accurate, the nature of what is being measured. In principle, years of formal education have no upper bound, but in practice very few people will acquire successive doctorates. Hence we might expect inequality in years of schooling to fall as average years of education rises. In the case of the achievement data, the item response modelling in principle removes any problem of achievement scores being capped by a theoretical maximum score, but we have insufficient knowledge of the technique to judge whether there is any problem in practice.

[9] Note that our results are conditional on both school enrolment and attendance. Children that are not enrolled or who do not attend schools cannot be tested. This may be a non-trivial issue for the poorer Eastern European countries in our 21 country pool.

How large are the within-country differences?

To this point we have not commented on the sizes of any of the within-country differences in achievement scores. How big are the inequalities in learning that we are measuring?

We need a metric for the achievement scores so that a given number of points can be interpreted in terms of something that is readily understood. The survey organisers provide what might be called partial metrics in the form of international benchmarks of achievement. These may be thought of as being similar to absolute international poverty lines, measured in, say, dollars per day.[10] For example, the PISA organisers define five levels of reading literacy. Children below level 2 are considered unable to 'locate straightforward information, make low-level inferences of various types, work out what a well-defined part of a text means and use some outside knowledge to understand it' (OECD 2001: 47). These thresholds are a useful guide but the metric they provide is partial. They tell one nothing directly about how to interpret any measure of the dispersion of scores, such as P95–P5 or the standard deviation. It is rather like having a tape measure to judge people's height that is blank except for a very few unevenly spaced marks, which are attached to labels describing something about height at that level in terms of what a person can or cannot do (e.g. 'see over others at a football game'). This measure can be used to find out the proportion of people with height at or above a given mark but it cannot be used to say something direct that compares the exact heights of two people.

Figure 2 suggests a way forward. The graphs show the distribution of science scores in the 1995 TIMSS round for a low inequality country, France, and for a country with above average inequality, Germany. The solid lines show the distributions for 8th grade pupils, which we use in Table 1. The dashed lines show the distributions for 7th grade pupils, who were also tested in TIMSS in 1995 and who faced the same test questions as the children in the 8th grade. The width of the distributions for the 8th grade children can be judged in terms of the change in average scores between 7th and 8th grades. In Germany, the value of P95–P5 is equal to 10.7 times the difference in these average scores. This is the highest for any country for which there are data on both grades, with the exception of the USA where the multiple is 13.3. But even in France, the country with the smallest value of P95–P5 in relation

[10] The analogy may not extend to dollars expressed in purchasing power terms, however. A given level of achievement is likely to have different implications from country to country.

to the difference in average scores across grades, the figure is equal to 5.5.[11] The mean across all countries of the difference in average scores between the two grades is an alternative yardstick. On this basis, the P95–P5 science values ranges from a multiple of 6.1 in Hong Kong to 8.6 in Romania. The analogous multiples for maths in TIMSS 1995 range from 7.2 in France to 10.2 in Romania. Similar sorts of figures (although including lower values) are found for P95–P5 for reading in PISA, applying the difference in average scores between grades attended by the 15 year olds in the survey.[12]

Viewed in this way, the differences within countries, including in low inequality countries, in a single school year or among children of the same age seem quite large everywhere. This should not come as a surprise, given evidence from differences in achievement that are recorded in national surveys or national exam results.

Inequality at the top and the bottom

The last aspect of inequality of learning that we consider are differences in the top half of the distribution compared to those in the bottom half. One might be more concerned with the latter, arguing that it is the extent to which those of lower ability fall short of the average level of achievement that should be the principal focus for policy on differences in educational opportunity.

Figure 3 shows each country's average rank on P95–P50 and on P50–P5 across the six tests in PISA, TIMSS and PIRLS. As in Table 1, the three surveys are weighted equally in these calculations. In general, larger inequalities in the top half of the distribution are certainly associated with larger inequalities in the bottom half. However, there is some variation about the 45 degree line. Countries above the line all do worse in rank terms on P95–P50 than on P50–P5: their levels of inequality in the top half of the distribution, compared to those of other countries, are larger (in rank terms) than their levels in the bottom half, again compared to other countries. It is noteworthy that all seven Central and Eastern European countries are in this category, although for several of these the average ranks on the top and bottom halves do not differ that much. Germany is perhaps the most obvious case below the line – Germany stands out more (in rank terms) for inequality in the bottom half of the distribution.

[11] The standard deviations for France and Germany, measured in the same way as multiples of the difference in average scores across grades, are 1.7 and 3.3 respectively. All the figures for TIMSS in this paragraph refer to the 1995 round.

[12] In unpublished work with Giorgina Brown and Robert Waldmann, we have further investigated a possible metric by looking at differences between 7th and 8th grade TIMSS scores at various points of the distributions.

The actual values of P95–P50 and P50–P5 for each of the six tests reveal that the distributions almost invariably display mild negative skew, with P50–P5 somewhat larger than P95–P50. (The differences are modest but can just be seen easy for Germany in Figure 2.) For reading in both PISA and PIRLS and for maths in PISA, this is true for all the 21 countries that we consider and there are only a handful of exceptions for the other tests.

Table 2 uncovers a related aspect of the distributions, showing the variation in the 21 country values of P5, P50 and P95, measured in points scores. In every test, the standard deviation of P5 exceeds that of P95, and in most cases that of P50 as well. Countries vary more on levels of achievement for people near the bottom of the distribution than they do for levels achieved by their high performers (given the particular item response models that have been used). The bottom half of the table shows this is also true for the subset of 14 OECD countries, although the differences in the standard deviations are in this case less marked.

4. Conclusions

Use of data from international surveys of learning achievement complements the analysis of differences in educational inequalities across countries that is based on attainment data. In order to avoid reliance on a single source, we have used data on six tests from three different surveys that relate to compulsory school age children, focusing on a pool of 21 countries covered in all the surveys. Our results show that (i) the surveys broadly agree on which countries in this pool have the greatest inequality in learning and which the least inequality, although care is often needed when drawing conclusions based on one source alone; (ii) lower inequality of learning and higher levels of average learning tend to be associated; (iii) differences in learning within all countries seem quite large in absolute terms; (iv) all surveys show more variation in low achievement across countries than in high achievement.

Future work in this area can head in several directions. First, there is the question of the extent to which differences in learning inequality across countries while at school help drive differences in inequality in earnings and other outcomes in later life. An important start to answering this question has been made by Bedard and Ferrall (2003), who compare inequality in achievement in international maths studies from 1964 and 1982 that were the forerunner of TIMSS, with earnings inequality recorded in survey data for the same population cohorts drawn from the Luxembourg Income Study. They find the two are positively associated. However, we feel they demand too much of the achievement data, since they make direct comparisons of the levels of inequality in achievement with the levels of earnings inequality. A complementary approach, adopted by Blau and Kahn (2005), is to use survey data that record both achievement scores and wages for the same individuals, although in this case the achievement scores do not relate to childhood, being collected at the same time as the wage data. Their study, based on the IALS data mentioned in the introduction, shows that greater inequality in cognitive skills in the USA than in other countries does help explain the higher US wage inequality, but only to a limited extent.

A second line of enquiry is to try to explain the observed differences in learning inequality in the international achievement surveys, resulting from family background, school institutions and combinations of the two (for example Schütz, Ursprung and Wößmann 2005, and Marks 2005). To date, such studies have been based on a single source and in line with our approach in this chapter we think that some comparisons of results across surveys is necessary. (A modest start is made in Micklewright and Schnepf 2004.) We think that it would be profitable to investigate the variation in results across different age groups. National studies based on panel data show inequalities in learning to be present at a very early age but to develop during childhood and the teenage years (for example Feinstein 2003, Carneiro and Heckman 2003). In the ideal world one would compare such studies across countries, but, in the absence of suitable data, the cross-sections relating to different age groups in the international achievement surveys are a good place to start.

Third, there is a need for more methodological work that investigates the sensitivity of results on within-country differences in achievement to choice of item response model, and that provides more guidance on how to interpret the data produced by these models. We have noted that the modelling process certainly can influence results on learning inequalities (which qualifies our substantive results summarised above) and we have tried to be cautious in our use and interpretation of the data. These are not data that can be treated in the same way as those on income or stature.

References

Atkinson, A. B. (1975). *The Economics of Inequality*. Oxford: Clarendon Press.

Atkinson, A. B., Cantillon, B., Marlier, E. and Nolan, B. (2002). *Social Indicators. The EU and Social Exclusion*. Oxford: Oxford University Press.

Beaton, A. (2000). 'The Importance of Item Response Theory (IRT) for Large Scale Assessments' in S. Carey (ed.), *Measuring Adult Literacy. The International Adult Literacy Survey (IALS) in the European Context*. London: Office for National Statistics.

Bedard, K. and Ferrall, C. (2003). 'Wage and Test Score Dispersion: Some International Evidence'. *Economics of Education Review*, 22: 31–43.

Blau, F. and Kahn, L. (2005). 'Do Cognitive Test Scores Explain Higher US Wage Inequality?'. *The Review of Economics and Statistics*, 87: 184–93.

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

Brown, G., Micklewright, J., Schnepf, S. V. and Waldmann, R. (2007). 'International Surveys of Educational Achievement: How Robust are the Findings?'. *Journal of the Royal Statistical Society*, Series A, forthcoming.

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

Carneiro, P. and Heckman, J. (2003). 'Human Capital Policy', in J. Heckman and A. Kruger (eds.), *Inequality in America: What Role for Human Capital Policies?* Cambridge MA: MIT Press.

Denny, K. (2003). 'New Methods for Comparing Literacy across Populations: Insights from the Measurement of Poverty'. *Journal of the Royal Statistical Society* Series A, 165: 481–93.

Feinstein, L. (2003). 'Inequality in the Early Cognitive Development of British Children in the 1970 Cohort'. *Economica*, 70: 73–98.

Marks, G. (2005). 'Cross-National Differences and Accounting for Social Class Inequalities in Education'. *International Sociology*, 20: 483–505.

Mayer, T. (1960). 'The Distribution of Ability and Earnings'. *The Review of Economics and Statistics*, 42: 189–95.

Micklewright, J. (1999). 'Education, Inequality and Transition'. *Economics of Transition*, 7: 342–76.

Micklewright, J. and Schnepf, S. V. (2004). 'Educational Achievement in English-Speaking Countries: Do Different Surveys Tell the Same Story?'. Discussion Paper 1186. Bonn: Institute for the Study of Labor (IZA).

Müller, W. (1996). ‘Class Inequalities in Educational Outcomes: Sweden in Comparative Perspective’, in R. Erikson and J. O. Jonsson (eds.), *Can Education be Equalised? The Swedish Case in Comparative Perspective*. Boulder CO: Westview Press.

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

Pradhan, M., Sahn, D. and Younger, S. (2003). ‘Decomposing World Health Inequality’. *Journal of Health Economics*, 22: 271–93.

Schütz, G., Ursprung, H. W. and Wößmann, L. (2005). ‘Education Policy and Equality of Opportunity’. Discussion Paper 1906. Bonn: Institute for the Study of Labor (IZA).

Shavit, Y. and Blossfeld, H.-P. (eds.) (1993). *Persistent Inequalities: A Comparative Study of Educational Attainment in Thirteen Countries*. Boulder CO: Westview Press.

Thomas, V., Wang, Y. and Fan, X. (2001). ‘Measuring Educational Inequality: Gini Coefficients of Education’. Policy Research Working Paper 2525. Washington: The World Bank.

UNICEF (United Nations Children’s Fund) (2001). *A Decade of Transition*. Regional Monitoring Report 6. Florence: UNICEF Innocenti Research Centre.

Wößmann, L. (2003). ‘Schooling Resources, Educational Institutions and Student Performance: the International Evidence’. *Oxford Bulletin of Economics and Statistics*, 65: 117–70.

Table 1: Z-scores for P95–P5 in PISA, TIMSS and PIRLS

	Average rank	Average z–score	PISA reading	PISA maths	PISA science	TIMSS maths	TIMSS science	PIRLS reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hong Kong	3.0	–1.2	–1.8	–0.4	–1.3	–1.1	–1.7	–1.0
Netherlands	3.8	–1.1	–1.2	–1.1	0.0	–0.9	–0.9	–1.5
Sweden	5.6	–0.8	–0.7	–0.4	–0.5	–1.4	–0.9	–0.7
France	6.5	–0.8	–0.8	–0.9	0.6	–1.8	–1.4	–0.4
Canada	7.3	–0.7	–0.4	–1.3	–1.0	–1.0	–0.8	–0.2
Iceland	7.5	–0.6	–0.8	–1.3	–1.2	–1.0	–0.5	0.0
Czech Republic	7.3	–0.5	–0.1	–0.1	–0.3	–0.2	–0.6	–0.9
Latvia	8.1	–0.5	0.6	0.4	0.2	–0.5	–0.8	–1.2
Hungary	10.1	–0.2	–0.6	–0.1	0.6	0.5	–0.2	–0.8
Italy	10.1	–0.2	–1.0	–0.7	0.1	0.6	0.2	–0.3
Norway	9.9	–0.2	0.8	–0.6	–0.2	–0.8	–1.2	0.5
Russia	11.0	0.0	–0.7	0.5	0.4	0.5	0.9	–0.8
Greece	12.9	0.1	0.0	0.9	0.0	0.6	–0.2	–0.2
Germany	12.5	0.1	1.9	0.4	0.7	–0.5	0.6	–0.7
UK	13.5	0.5	0.4	–0.6	0.2	0.2	0.6	1.0
Bulgaria	14.9	0.6	0.6	1.0	–0.1	0.6	0.8	0.6
USA	16.8	0.7	1.2	0.0	0.4	0.7	1.2	0.6
Macedonia	16.0	0.9	–0.7	0.0	–1.7	1.4	1.2	2.1
New Zealand	17.1	0.9	1.4	0.0	0.3	0.9	0.7	1.5
Romania	16.7	1.0	0.5	1.5	–0.3	1.3	1.1	1.2
Israel	20.6	2.0	1.6	2.9	3.3	1.7	2.0	1.5

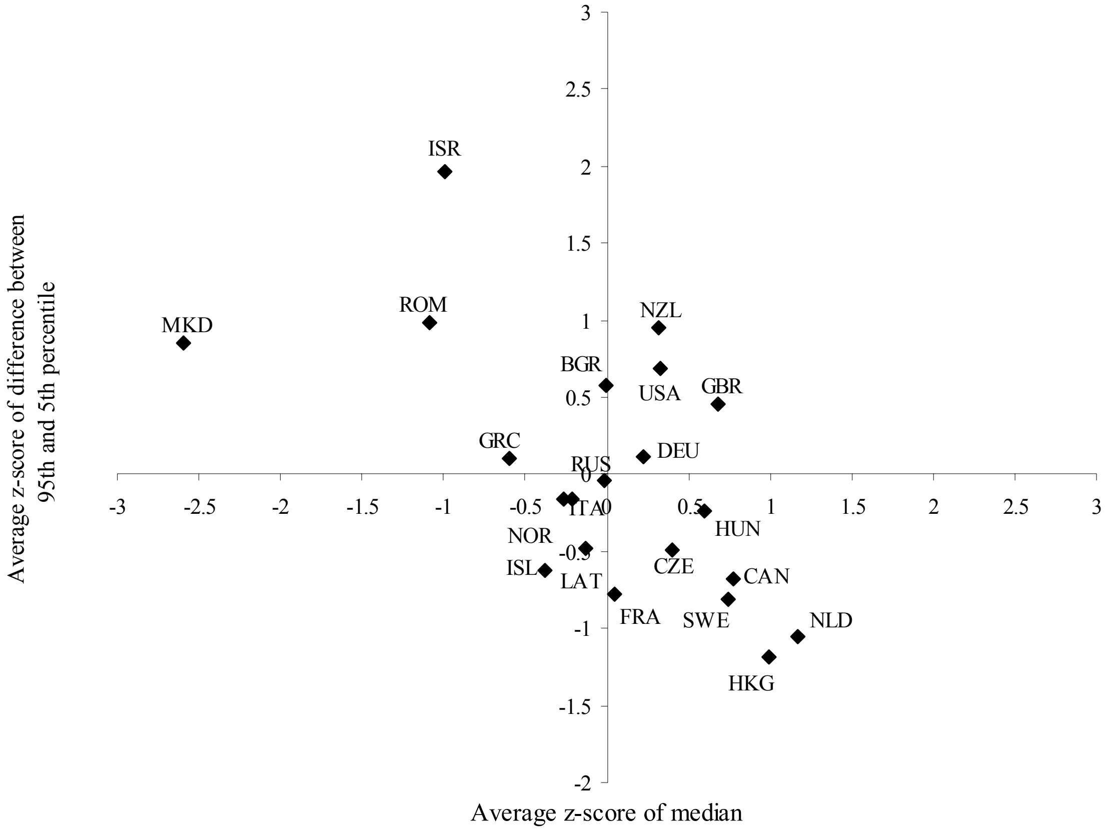
Note: countries are ordered by average z-score. The z-score adjustment is described in the text. A negative z-score implies a lower difference between P95 and P5 than the average for the 21 countries, a positive z-score implies a larger difference than the average.

Table 2: Standard deviations of country values of selected score percentiles

	PISA reading	PISA maths	PISA science	TIMSS maths	TIMSS science	PIRLS read
All 21 countries						
P5	42.5	59.8	39.2	40.9	38.4	44.0
P50	42.5	48.1	39.5	29.6	26.0	24.6
P95	38.4	36.4	34.8	24.3	25.7	20.1
14 OECD members						
P5	24.6	39.1	23.1	28.7	27.3	30.0
P50	20.1	31.5	21.9	20.6	23.0	16.0
P95	21.6	24.9	19.9	19.1	25.9	17.3

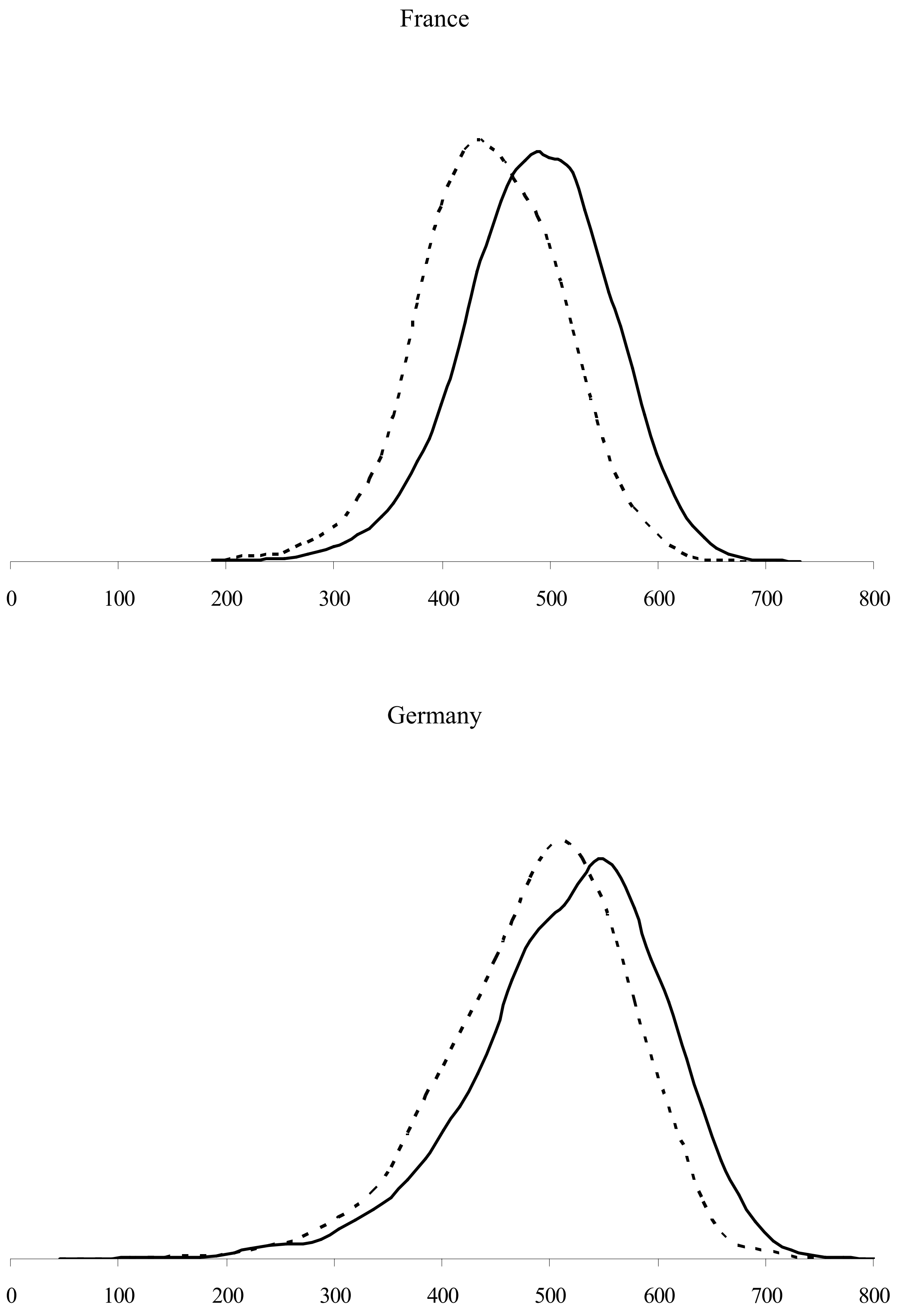
Note: the standard deviations relate to the scores in each survey unadjusted by any z-score transformation.

Figure 1: Average z-scores for P95–P5 and P50 in six tests in PISA, TIMSS and PIRLS



Note: the z-score adjustment is described in the text. The values plotted are the average z-scores across the six tests (weighting surveys equally). The correlation between the two measures is -0.55.

Figure 2: Distributions of TIMSS science scores

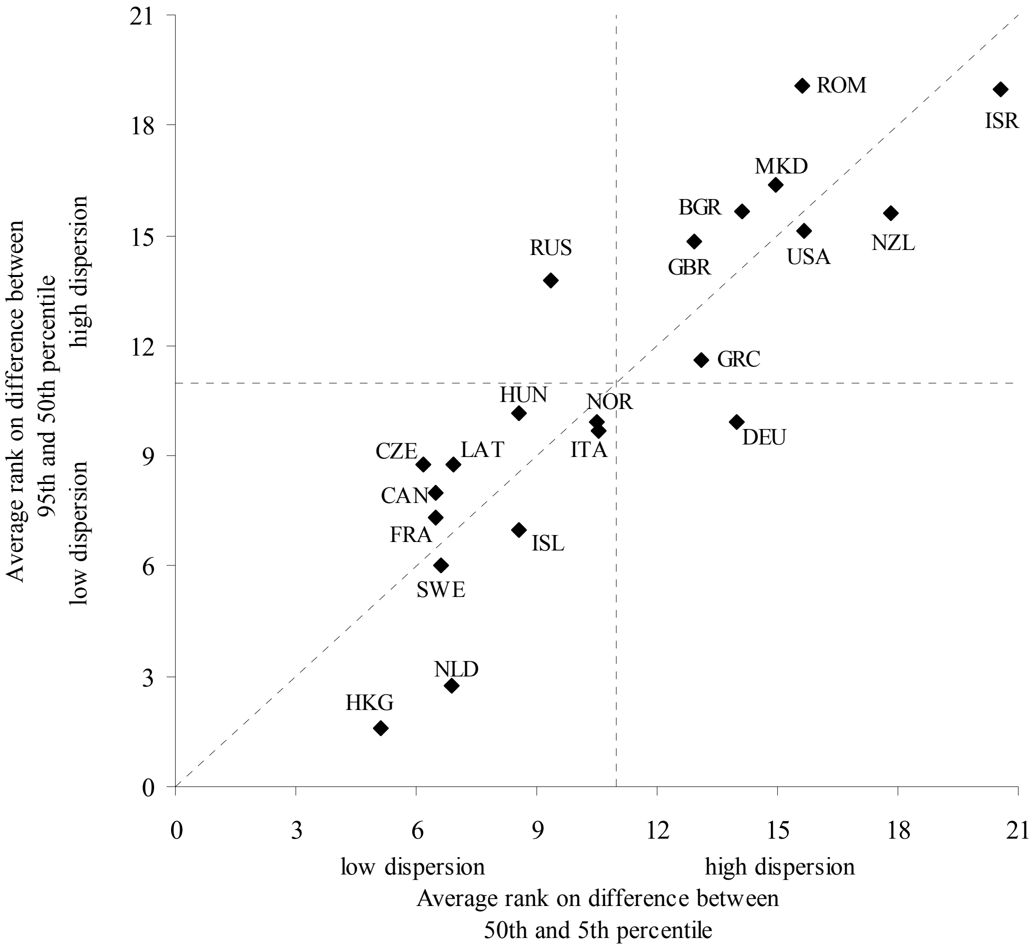


France

Germany

Note: the solid lines show the distributions for the 8th grade, the dashed lines show the distributions for the 7th grade.

Figure 3: Average rank on P95–P50 and P50–P5 in 6 tests in PISA, TIMSS and PIRLS



Note: higher values of the average ranks indicate indicate larger differences between the quantiles in question.