USING INTERNATIONAL SURVEYS OF ACHIEVEMENT AND LITERACY: A VIEW FROM THE OUTSIDE

By Giorgina Brown[1] and John Micklewright[2]

UNESCO Institute for Statistics, Montreal, 2004

[1] ISTAT, Rome

[2] School of Social Sciences and S3RI, University of Southampton

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946. The Organization currently has 190 Member States.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law, and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfill its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical co-operation to Member States for their development policies and projects; and 5) the exchange of specialized information.

UNESCO is headquartered in Paris, France.

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science and technology, culture and communication.

UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

UIS is based in Montreal, Canada.

UNESCO Institute for Statistics
P.O. Box 6128, Succursale Centre-Ville
Montreal, Quebec H3C 3J7
Canada

Tel: (1 514) 343-6880
Fax: (1 514) 343-6882
Email: uis@unesco.org
http://www.uis.unesco.org

ISBN 92-9189-013-8

Ref: UIS/TD/04-02

The paper reports on experience of using three international surveys of learning achievement or functional literacy: TIMSS, PISA and IALS. The continued development of new surveys means that it is all the more important for users to share their knowledge of the sources. A range of issues are considered of which the most important are (i) the robustness of results to the method used for aggregating the answers for each individual into a single score; (ii) the methods that can be used for presenting summary statistics; and (iii) the need to systematically compare the basic results of the different surveys.

Acknowledgements

This paper was financed by a research grant from UNESCO Institute for Statistics, Montreal. We acknowledge joint unpublished work with Robert Waldmann, University of Tor Vergata, Rome, and the assistance of Sylke Schnepf, University of Southampton. We have benefited from discussions with the organizers of TIMSS and PISA (Michael Mills, Ina Mullis, Eugene Gonzalez and Andreas Schleicher), and we are grateful to them for their help but they are not responsible for the way we have represented their data. The views expressed in this paper are our own and should not be associated with UNESCO, ISTAT or the University of Southampton.

Table of Contents

Page

1. Introduction 7

2. Survey content and the nature of the data 8
   a) Main characteristics and differences in target populations and coverage 8
   b) The survey reports 12
   c) The microdata 14

3. The 'scaling' of the scores and the calculation of standard errors 15
   a) Item response models and derivation of summary scores 15
   b) The calculation of standard errors 25
   c) The standard error of a difference in quantiles and the Bonferroni adjustment 29

4. Methods for presenting summary results 33
   a) Benchmarks and levels – 'absolute' educational disadvantage 33
   b) Distribution of achievement 35
   c) Scattergrams of dispersion versus central tendency 36
   d) 'Relative' educational disadvantage 38
   e) Graphical representations: Kernel density plots – and the search for a 'metric' 40

5. Comparing the results of the different surveys 43
   a) Traffic lights 43
   b) Average ranks 46

6. Conclusions 48

7. Country and region abbreviations 49

8. List of Figures and Tables 50

9. References 51

10. Appendix A: Recent and forthcoming cross-national surveys of learning achievement and functional literacy 52

11. Appendix B: Which countries participated in which surveys? 53

1. Introduction

Recent years have seen an increasing number of international surveys of learning achievement of children and of functional literacy of adults. These include the International Adult Literacy Survey (IALS), the Trends in International Maths and Science Study (TIMSS), the OECD Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS). The next few years will see the publication of results from further rounds of TIMSS and PISA, along with the successor to IALS, the Adult Literacy and Life Skills Survey (ALLS). Details are given in **Appendix A**, and the countries included in the surveys from which results have already been published are listed in **Appendix B**.

This paper will draw on experience from using and interpreting data from TIMSS, PISA and IALS. A range of practical issues of importance to users and potential users of these surveys will be discussed. The research community is now faced by a plethora of survey data on achievement/literacy. It is important that lessons learned about the quirks and peculiarities of each source and about appropriate methods for analysis and presentation of this sort of data are documented and made widely available. The paper will cover use of both published summary data and microdata from the surveys.

Our aim is not to provide a rounded 'user guide' that would substitute for a reading of each survey's own documentation. Rather it is to document hands-on experience of various features of the surveys obtained by a team of users from *outside* the organizations collecting the data, complementing rather than substituting those organizations' own guides to their data. Much of the experience that the paper will document was obtained through our work during 2001/2 at UNICEF Innocenti Research Centre, Florence on a report on educational disadvantage in OECD countries (UNICEF 2002). Our work in this report focused on comparisons of basic results on levels of achievement/literacy in the different surveys, and it is our experience in this area of analysis that is the subject of the current paper. We certainly do not claim experience of all or even most aspects of what are very rich data sets providing information on many different aspects of the knowledge generation process.

Section 2 briefly describes the content and nature of TIMSS, PISA and IALS and what unit record microdata (along with sample computer programmes for their analysis) are available from the survey organizers for secondary analysis.

Section 3 deals with two issues that any user has to come to terms with very quickly at some level. The first is the psychometric 'scaling' that is used by the survey organizers to summarise the respondents' answers into a single score. We show how the published results can be sensitive to the precise scaling method that is chosen. The second is the calculation of standard errors for summary statistics in the light of the complex sample design of the surveys.

Section 4 deals with methods for presenting summary results, a key issue given that the achievement score data have no natural 'metric' or measuring rod. We start by discussing the methods used by the survey organizers, including 'benchmark' levels taken to represent a given level of ability and graphical representation of measures of central tendency and dispersion. We then describe – and justify – some alternatives that we have found useful in our own work with the data, including other representations of central tendency versus dispersion and kernel density plots of the distribution of scores.

Section 5 discusses the direct comparison of results from the different surveys, something that we feel there has been too little of to date. What seem to be appropriate methods to see how the surveys compare on country-level measures of central tendency and dispersion? We also describe one method we have found useful for combining results of the different surveys into one summary statistic.

Section 6 presents concluding remarks on the paper's analysis.

2. Survey content and the nature of the data

a) Main characteristics and differences in target populations and coverage

In order to appreciate differences between the three surveys, which is essential before we can look at and compare the results, we now give a brief review of each survey with a focus on their distinctive features.

TIMSS was conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA), which has been carrying out international studies of school achievement, attitudes and curricula since 1959. It intended "to provide a base from which policy-makers, curriculum specialists and researchers could better understand the performance of their educational systems" and to determine the extent to which pupils could understand and apply essential maths and science knowledge.

In TIMSS, the target populations studied in 1995 were children in the two grades in which most 9-year-olds (3rd and 4th grade) and 13-year-olds (7th and 8th grade) were enrolled and children in the last grade of secondary school. In 1999, the target population was children in the higher of the two grades in which most 13-year-olds were enrolled (the average age of these children across participating countries was 14.4 years). Conventionally, this grade is referred to as the 8th grade, since in most countries it refers to the eighth year of formal schooling, but for example students in Denmark, Finland, Norway and Sweden had one year less of formal schooling, while students in New Zealand and the United Kingdom had one year more. Between 1995 and 1999, 52 countries, including 27 OECD members, participated in the study in one or other – or both – years.

Here we focus on 8th-grade results, pooling countries which participated in 1995 and 1999. For countries which participated in both 1995 and 1999, we use the 1999 data. For those that participated only in 1995, we use the 're-scaled' version of the data, i.e. data in which the method of psychometric 'scaling' is the same as that applied to the 1999 data (*see Section 3 below*). However, we also refer to the 3rd, 4th and 7th grades in 1995 to establish a metric, since 3rd- and 4th-graders and 7th- and 8th-graders completed the same tests (*see Section 5 below*). About one-third of the questions to 8th-graders in 1999 were exactly the same as those put to 7th- and 8th-graders in 1995. The others were intended to give results that were comparable.

The focus on a *grade* rather than on children of a certain *age* (as in PISA) is worth noting. 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 achieved before passage upwards is allowed. Countries in the latter group should have a shorter lower tail of achievement, other things being equal, and this could explain some of the differences in the results between TIMSS and PISA.

TIMSS collected data through first randomly sampling schools – about 150 per country – and then testing *all children in one class* drawn at random from each sampled school. The procedure is thus essentially one of two-stage cluster sampling. The first stage sample is stratified, and in a small minority of countries, there is in fact a third stage with sampling of students within classes.

Samples consisted on average of 3,800 8th-grade children per OECD country, with an average overall response rate of 88% after replacement of non-responding schools with substitutes. But non-response among schools was a serious problem in some countries, if we look at rates before replacement (on average 81%). The worst cases are Austria and England. Without allowance for replacement, the overall response rates in these countries slip to less than 50%. (In contrast, non-response among children within responding schools is low everywhere; the average student response rate is 94%).

It seems possible that schools that do not initially respond are those where achievement is lower than the national average, the schools fearing that they will be shown in a bad light by the survey's results. The replacement procedure involved selecting, *a priori*, two replacement schools for each sampled school. The TIMSS 1999 technical report argues that the use of stratification in the sample design and the ordering of the school sampling frame by size "ensured that any sampled school's replacement would have similar characteristics" (Martin et al, 2000b, p.38) and that this minimises potential bias.

Response rates are calculated in relation to the target population. TIMSS countries were allowed to restrict the latter in two ways. First, certain types of schools (or children within schools) could be excluded (less than 10%) because it would prove to be very difficult or expensive to test, e.g. schools for students with special needs and small remote rural schools. Children could also be excluded if they had received less than one year of instruction in the language of the test (i.e. some non-native language speakers). The response rates are calculated on the potential samples net of these exclusions. Exclusions among all OECD countries participating in 1995 and 1999 averaged 4% of the potential sample but reached as high as 9-10% in Spain and Germany. Second, countries could exclude whole regions or parts of their educational system where it was not possible to test the entire target population. Among the OECD countries, this occurred only in Germany (12%) and Switzerland (14%). One of the 16 German regions (Baden-Wuerttenberg) and 4 out of the 26 Swiss cantons declined to take part in TIMSS.

Apart from the achievement tests in maths and science, children, teachers and school principals responded to questionnaires collecting a variety of information on individual background and the context of learning. See examples of test questions from each TIMSS, PISA and IALS in UNICEF (2002, p.10). TIMSS has more multiple-choice questions – approximately two-thirds of the items in 1999 – than PISA, while IALS has no multiple-choice questions at all.

As do PISA and IALS, TIMSS provides an estimate of each child's achievement in the form of a summary score based on the application of "item response models" to the answers given to each question. This process is also known as "scaling". The 1995 and 1999 rounds of TIMSS used different procedures to scale the data, and the results in the published reports for each of these two rounds are therefore not comparable. However, the 1995 data were re-scaled by the International Study Center at Boston College using the same model as in 1999[1] (*see Section 3 below*).

PISA intends "to measure how well 15-year-olds, approaching the end of compulsory schooling, are prepared to meet the challenges of today's societies" on the basis of their ability in reading, scientific and mathematical literacy. PISA has chosen a more ambitious path in the attempt 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". While TIMSS focuses more on measuring mastery of an internationally agreed curriculum, PISA is intended to measure broader skills, trying to look at how students would be able to use what they have learned in real-life situations. But it is hard to pinpoint what this entails in practice.

Apart from the achievement tests, children and school principals in PISA responded to questionnaires on family background and the school respectively (differently from TIMSS, the children's teachers were not administered a questionnaire).

---

[1] We use the re-scaled 1995 data when we pool countries from 1995 and 1999.

First administered in 2000 in 32 mainly-OECD countries, with a main focus on reading, PISA was extended in 2001 to an additional 11 countries ('PISA plus'). In this paper we use data from the 2000 round of PISA only. Assessments will subsequently occur every three years (with the main focus on maths in 2003 and on science in 2006).

The target population for PISA consists of all 15-year-olds in school irrespective of the grade they are in. On average, 15-year-olds in the OECD have been attending school for between 8.9 years (Finland, Switzerland) and 11 years (New Zealand). The sampling design used in most countries was a two-stage stratified sample[2]. First, schools were selected from a stratified sample, and then, 35 15-year-old students within the sampled schools were selected at random (or all 15-year-olds if fewer than 35 were in the school).

This is a key difference compared to TIMSS, which selected an entire class at random and tested all students in that class. This makes TIMSS more suitable for measuring peer or teacher effects within *one classroom*. On the other hand, PISA provides a random sample of all students (of one age) *in the school*, which has its own attractions. The two sampling procedures will clearly produce different types of samples in terms of student achievement in the situation where schools stream students by ability so that students in one class are not representative of those of the same age in other classes of the same school. One result of such streaming is that while PISA can provide an estimate of the within-school and between-school division of variance of achievement, TIMSS cannot.

In PISA, an average of almost 5,700 15-year-olds were assessed per OECD country, with an average overall response rate of 85% after replacement of non-responding schools. Here, too, the shortfall comes at school level (average 86% before replacement and 92% after replacement) or individual student level (average 90% after replacement).

As in TIMSS, some schools could be excluded from the target population, for example schools in remote, inaccessible areas. Also mentally retarded and functionally disabled students, as well as non-native language speakers, could be excluded. It was required that the overall exclusion rate within a country be kept below 5% (the average was 3%), although the rate in Luxembourg was 9% and in Poland 9.7% (in the latter mostly due to 15-year-olds in primary schools being excluded). In the United Kingdom the region of Wales was excluded from participation, while in Belgium, this was the case with the German-speaking community.

[2] In three countries, a three-stage design was used: geographical areas were sampled first.

Exclusions and response rates were carefully analysed for each country, and stratification and weighting adjustments were used in such a way as to minimise any bias which might have been introduced. The only country which did not reach PISA standards was the Netherlands (only 27% of school participation before replacement and 56% after replacement) and was thereby excluded from most country comparisons in the international report. After analysis of potential non-response bias, data for the United Kingdom and the United States was considered acceptable even with low school response rates (respectively 61% and 56% before replacement and 82% and 70% after replacement).

In IALS, different countries participated in the three different rounds of data collection, with a total of 21 countries (mostly OECD). IALS was designed to measure the extent to which people of working age (16 to 65) are able to use literacy skills to perform everyday tasks, through the assessment of proficiency in three areas: prose literacy (understanding and using information from texts), document literacy (locating and using information contained in various formats) and quantitative literacy (applying arithmetic operations to numbers embedded in printed material). Literacy is defined as "the ability to understand and employ printed information in daily activities, at home, at work and in the community – to achieve one's goals, and to develop one's knowledge and potential".

Samples (nationally representative of the adult population aged 16 to 65) averaged 3,400 persons per country, including nearly 700 young people aged 16 to 25. The survey was conducted in people's homes with an average response rate of 62%, so the nature of the sampling process was very different to that in TIMSS and PISA: in contrast to those surveys, IALS is a household survey. A background questionnaire collected information on a variety of subjects, including labour market activity.

b) The survey reports

For TIMSS, two separate reports for the maths and science assessments were published by the survey organizers (for the 1999 data collection, Mullis et al., 2000 and Martin et al., 2000a). These can be downloaded or bought from www.timss.org. These reports tend to describe achievement according to international benchmarks and content areas, and in relationship to several student background and attitude variables, teacher instruction methods, school and curriculum contexts taken separately. Other specific reports are also available, including two reports on a voluntary benchmarking study in which 27 separate jurisdictions of the United States (13 states and 14 districts or consortia) participated, each with its own separate sample, in order to compare themselves to the United States as a whole and to other TIMSS 1999 countries. (The sample sizes in each of these 27 jurisdictions are comparable in size to the country samples in the main TIMSS surveys.)

PISA produced one main report, including results from the reading, maths and science assessments (OECD, 2001). A second report including 'PISA plus' countries was published in 2003 (OECD and UNESCO Institute for Statistics, 2003). These can be downloaded or bought from www.pisa.oecd.org. Specific thematic reports are also being published. The main report, as well as showing a profile of student performance and differences according to student background and learning environment, tries to explain the differences in student performance and to give some indications for policy (using multivariate methods, among others). The PISA reports place more emphasis on family background factors in their analysis of the data than do the TIMSS reports. One aspect of this is the use of indices based on principal component analysis to summarise family background variables.

IALS reports were published after each round of data collection – the last one (OECD and Statistics Canada, 2000) includes countries from all three rounds except Italy, which has a separate Italian report (Gallina, 2000). The IALS international report can be bought from the online OECD bookshop (www.oecd.org: a browsable but not printable version is available online). Information on IALS is also available on www.nald.ca/nls/ials/introduc.htm. Information on the successor of IALS, the Adult Literacy and Lifeskills Survey (ALLS) is available on www.ets.org/all/index.html. Apart from looking at the distribution of literacy skills in the population, the main IALS report looks at context and relationships with educational attainment, participation in adult education and other activities, and outcomes such as occupation and earnings.

How to replicate results as published in the survey reports? It seems to us that typically there is not enough guidance in the reports on this matter. To take one example, in all three surveys, in order to provide consistent population estimates (for example the country mean), a set of five "plausible values" for each individual are selected at random from an estimated ability distribution of scores that could be reasonably assigned to each individual (*see Section 3 for details*). Use of all plausible values is necessary for calculating the measurement (or imputation) error component of the standard error. But for the estimates themselves, it seems that one can use alternatively a single plausible value or the average of all five. In fact, in the TIMSS 1995 reports, the first plausible value only is used for the population estimates, while in the TIMSS 1999 reports, the average of the five is used (as in PISA). Note that to replicate the percentiles in the report, one must not average the five values first and then calculate the percentiles, but calculate the percentiles five times separately and then average them[3]. In the IALS report, some tables were calculated averaging all five plausible values while others used only the first.

[3] The 50th percentile (or median) of achievement is included in the TIMSS reports but not in the PISA or IALS ones.

c) The microdata

Microdata for TIMSS and PISA are available on their respective websites, mentioned above. Microdata for IALS cannot be downloaded from the web but are available on CD-Rom from Statistics Canada (www.statcan.ca). The Australian data are not included in the IALS CD-Rom, but specific data requests can be made to the Australian Bureau of Statistics, National Centre for Education and Training Statistics (www.abs.gov.au). (The Italian microdata however are included, despite Italian data being absent from the international report – see above.) All these files contain responses to test and questionnaire items, the "plausible values", weighting variables and some constructed indices.

TIMSS provides a separate file for each country participating in the survey. A CD-Rom can be bought from TIMSS to avoid the lengthy procedure of downloading them all. Programmes for calculating standard errors using the jackknife method in SPSS and SAS are included[4], as well as the data files, data almanacs with summary statistics and codebook files. Both TIMSS and PISA provide control files to convert the text data files into SAS or SPSS files. A Technical Report and User Guide is provided.

In PISA, most indices or derived variables used in the report are included in the microdata files, but not the "index of economic, social and cultural status", which is rather surprising given that it was the main index used in the concluding chapter of the international report (Chapter 8 on "What makes a difference to PISA results: some indications for policy"). A Technical Report and a Database Manual are available, as well as codebook and questionnaire files, and a compendium of the results by variable. The PISA database manual contains a description of what to do in order to calculate standard errors in the statistical packages SPSS, SAS or WesVar, but the programmes themselves are not provided. On the PISA website, as an alternative to downloading the microdata files, one can select certain variables and receive immediately online the tables containing means, standard errors and percentage distributions. There is also the possibility of submitting a multi-dimensional query to an automated service. Furthermore, there is a PISA data service helpdesk which, for a fee, answers specific customised queries.

The IALS User Manual contains print-outs of programmes in SPSS and SAS for calculating standard errors. But in general it is less detailed than the TIMSS and PISA manuals. Questionnaires and record layouts are included.

---

[4] These are needed to calculate the sampling variance in a way which takes account of the clustering design and to estimate the imputation variance.

STATA users are left to themselves by the organizers of all three surveys, which is unfortunate given the spread in recent years of this package in the social sciences. Section 3 deals more extensively with ways in which standard errors can be calculated.

3. The 'scaling' of the scores and the calculation of standard errors

a) *Item response models and derivation of summary scores*

The answers that a respondent gives to the questions in the surveys are summarised by the organizers into a single score for the subject concerned – maths, science, reading, different types of literacy, etc. The aggregation of test answers into a single score is a complex process and a far cry from simple procedures like counting the number of correct answers. We feel that the descriptions of the procedure in the survey reports are typically not very transparent and that, in particular, they are not written in ways designed to be understood by those who are unfamiliar with the literature on psychometrics, the science of psychological measurement/testing.

The scores are produced using 'item response' (IR) models. (Beaton, 2000) provides an accessible account of the arguments for use of these models rather than simpler methods of summarising respondents' answers and we do not enter that debate here.) The scale for the scores is chosen by the survey organizers, typically so that the international mean is 500 and the international standard deviation is 100 (the mean and standard deviation in the data pooled for all participating countries). This gives the name sometimes given to the procedure, 'scaling', that we use here.

Two issues arise. First, we feel that a somewhat simpler and more intuitive explanation is needed by many survey users of how the scaling process works. While the detailed descriptions that are given in the survey reports are appropriate for those with a good knowledge of psychometrics or other readers with an advanced statistical training, we think the standard user needs more intuition in order to understand what is going on. We try to supply some of this below (although the level of technicality remains higher than some readers may feel comfortable with).

Second, the user is not in general told in the survey reports whether key results, e.g. the main summary statistics, are robust or not to the choices that are made in the scaling procedure. This contrasts with what is now established practice in some other disciplines, for example the use of econometrics within labour economics (e.g. Mroz, 1987) or the analysis of household income survey data on poverty (e.g. Atkinson, 1998).

And the procedure is sufficiently complex that it is utterly impractical for the vast majority of users making secondary analysis of the microdata to estimate IR models for themselves so as to gauge the sensitivity of results to alternative methods. As a step in what we see as the right direction, we show the sensitivity of TIMSS results from 1995 to two different methods of scaling. The basic idea here is that achievement scores are *derived* variables and the issue naturally arises as to whether the method of derivation has any appreciable impact on results that can be obtained with the data.

We use the example of TIMSS to illustrate the process of scaling. The following description is not an exact account (and simplifies deliberately in some places) but is intended to give the flavour of what is going on.

Let the probability of a correct answer to question *i* by student *j* be given by the logit function:

$$p_{ij}\,(\text{correct answer}) = 1/(1+\exp[-(\theta_j - \alpha_i)] \qquad (1)$$

where θ_j stands for a student's 'proficiency' in the subject and α_i denotes the difficulty of the question, allowing for the fact that some questions are easier than others and some are harder. Both θ_j and α_i are unobserved parameters to be estimated. Equation (1) represents a 'one parameter' IR model (often known as a Rasch model), meaning that there is just one parameter, α_i, relating to the question. We describe an alternative 'three parameter' model later.

The first step is to estimate the α_i by maximising a likelihood that has been conditioned on sufficient statistics for the θ_j, implying that the latter are treated as unobserved fixed effects at this point. A 'sufficient statistic' for θ_j is the number of correct answers by student *j*. In other words, an expression for the probability of observing a correct answer to question *i* by person *j* is written down that is made *conditional* on observing the total number of correct answers by person *j*. The expression for this conditional probability does not contain θ_j. Or, to put it another way, it appears in both numerator and denominator of the conditional probability in a way that allows it to be cancelled out of the expression. (It is this property that means that the number of correct answers by student *j* is a 'sufficient statistic' for the θ_j.) This means that the likelihood can be maximised with respect to the α_i alone without worrying about the values of θ_j . Estimation at this stage uses a sub-sample of the data for each country, pooled into one sample.

With estimates of the α_i in the bag, the second step in the process treats θ_j as a normally distributed random effect. The mean and variance of θ are then estimated separately for each country (using the country's full data set), treating the α_i as given. In principle the whole process could stop here: after all, an estimate of central tendency and dispersion in achievement for each country has been obtained.

But this would be rather unattractive – for example normality would have been forced onto the score distribution and no particular score could be attributed to each individual in the data set (all one would have is estimates of two summary statistics for each country, mean and variance).

The third step is to form a likelihood for p(θ | exam script of *j*), that is to write down a function for each individual *j* of the probability of θ given individual *j*'s answers to the TIMSS test. This likelihood (technically a 'posterior probability') is as follows:

$$p(\theta \mid \text{script}_j) = p(\theta \cap \text{script}_j) / p(\text{script}_j) \qquad (2)$$

$$= [p(\text{script}_j \mid \theta_j) \,.\, p(\theta)] / p(\text{script}_j) \qquad (3)$$

The first term in the square brackets in (3), $p(\text{script}_j \mid \theta)$, is a product of the expression in (1) for each question correctly answered and of one minus that expression for each question incorrectly answered. The second term, $p(\theta)$, is provided by the second stage in the estimation process described above – this is just the standard normal probability density function. The denominator, $p(\text{script}_j)$, is the integral of the numerator over all values from $-\infty$ to $+\infty$. Hence (3) is a (complex) function of data and parameters already estimated at the first and second steps.

The final step is to generate a random variable for each individual with the same probability density function as (3) and to draw five times from its distribution at random. These five numbers are known as 'plausible values', and should be interpreted as 'random numbers drawn from the distribution of scores that could reasonably be assigned to each individual' (Adams and Wu 2002). These five numbers can then be averaged to give a single number that serves as an estimate of the score for an individual, although as we note later one would not always want to average in this way (*see also Section 2*).

Broadly speaking, the procedure just described is that followed by TIMSS in 1995 to produce what we call the 'old scale' TIMSS scores. The 1999 procedure, resulting in the 'new scale' (or 're-scaled') scores, differed in two important respects. First, the logit function for the probability for a correct answer had two additional parameters relating to the questions, making three in total, hence a 'three parameter model'

$$p_{ij}(\text{correct answer}) = \gamma_i + (1 - \gamma_i) / [1 + \exp(-\beta_i(\theta_j - \alpha_i))] \qquad (4)$$

where γ_i is the probability that the answer to question *i* is guessed and β_i measures the power of a question to discriminate between individuals of high and low ability.

The other change was that at the second stage, θ is modelled as a function of observable characteristics of the student and his or her school, including a vast array of indices derived from principal component analysis of family and school background variables. A three parameter model along these lines was also then applied retrospectively by the TIMSS organizers to the 1995 data so results on both bases are available for that year.

A systematic comparison of the 'old' and 'new' scale results does not seem to be publicly available. In what follows we undertake some comparison, restricting ourselves to summary statistics of the sets of scores derived from the 1995 data using the one-parameter and the three-parameter IR models.

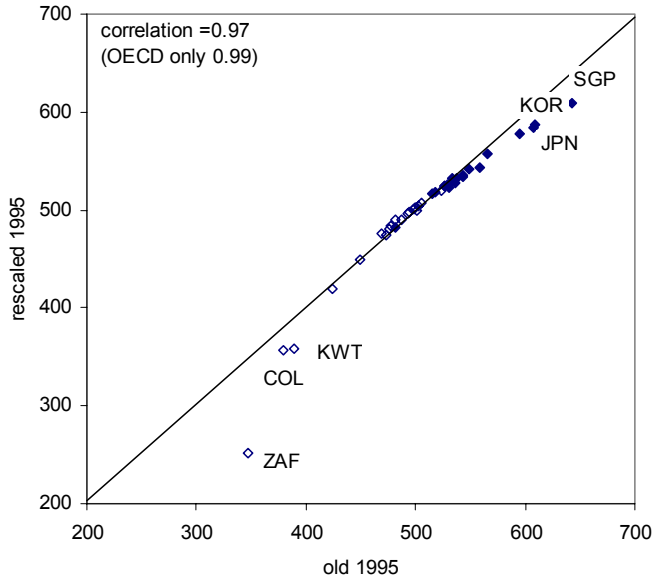
In the 1999 survey's technical report the TIMSS organizers argued that direct comparison of old and new scale scores is not appropriate because the new scale scores are based on parameters estimated with 8th grade students only whereas the old scale parameters were estimated with both 7th and 8th grade students (Yamamoto and Kulick, 2000: 253). (TIMSS 1995 covered both grades while TIMSS 1999 covered 8th grade students only – *see Table A*.) This implies that the mean new scale scores for 8th graders in the 1995 data are slightly lower than the old scale scores. This is because on the old scale, the 'base' was 7th and 8th graders taken together, against which any individual 8th grader would appear to score more highly than against a base of 8th graders alone. And the variances are slightly higher compared to those in the old scale scores.

However, we do not believe this difference in old and new scale scores noted by the TIMSS organizers invalidates a comparison of the *overall cross-country pattern of central tendency and dispersion* in the two sets of results.[5] Is the ranking of countries by mean achievement the same with old and new scale scores? Are conclusions about which countries have the highest variance in achievement robust to the scaling method? It is this form of comparison that we now make.

We start first with central tendency, as measured by the median. **Figure 1** plots the median for maths 'new scale' (1999 method) scores against the median for maths 'old scale' scores (1995 method) for all those countries in TIMSS in 1995. To be clear: the underlying raw data – the answers given by respondents to the questions – are identical in the two sets of scores. What differs is the method used to aggregate those data for each individual into a single score.

[5] For example, lower mean values with the new scale scores would be consistent with a correlation coefficient of 1.0 between old and new scale means provided every mean had changed in a way that could be described by the linear relationship NEW = a + b.OLD.

Figure 1: Old v. rescaled Q50, TIMSS 1995 8th-grade maths



The conclusion seems straightforward. The medians are very highly correlated indeed, both among just the OECD countries present in TIMSS 1995 and among all countries covered by the survey in that year.

The open-diamond countries in this and subsequent diagrams are those in the bottom half of the distribution of q5 values (the 5th percentile) on the old scale scores. A few countries lie some way off the 45 degree line, with South Africa (ZAF) being the most extreme case with a fall in median from old to new scale of over 75 points – a big difference. But the change in the scaling procedure hardly changes one's view of the ranking in a 'league table' of countries' levels of mean achievement. The same results are found for science (*not shown*).

We now turn to dispersion, as measured by the difference between 95th and 5th percentiles for maths, q95-q5. (This measure of dispersion is very highly correlated in the data with both the interquartile range, q75-q25, and the standard deviation – the results that follow are not driven by our particular choice of dispersion measure.)[6] As a preliminary, we first look at what happens to q5 and q95 separately in the re-scaling, shown in **Figures 2** and **3**.

[6] For example, among OECD countries the country standard deviations for maths have a correlation of 0.97 with q75-q25 and 0.99 with q95-q5.

Figure 2: Old v. rescaled Q5, TIMSS 1995 8th-grade maths

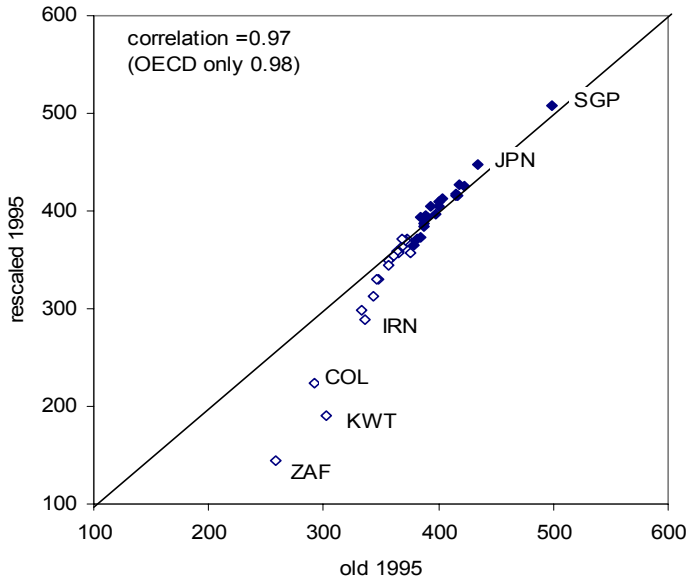
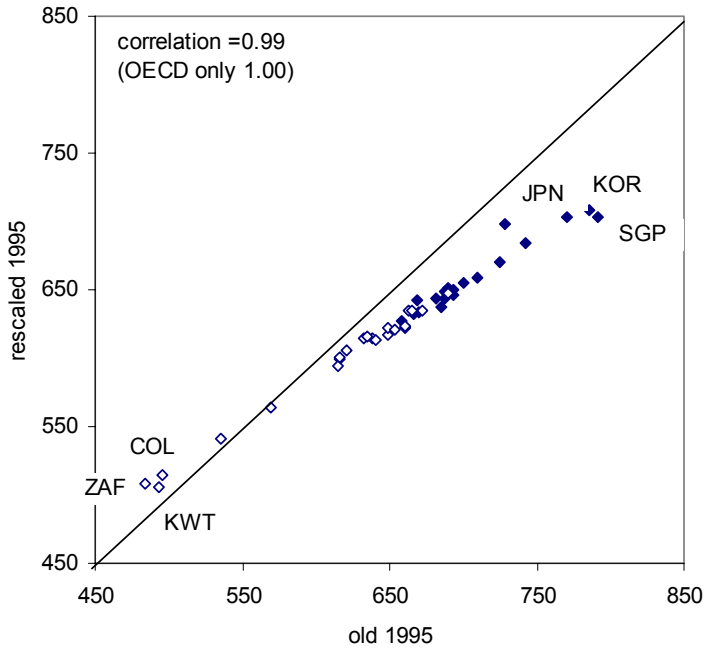


Figure 3: Old v. rescaled Q95, TIMSS 1995 8th-grade maths



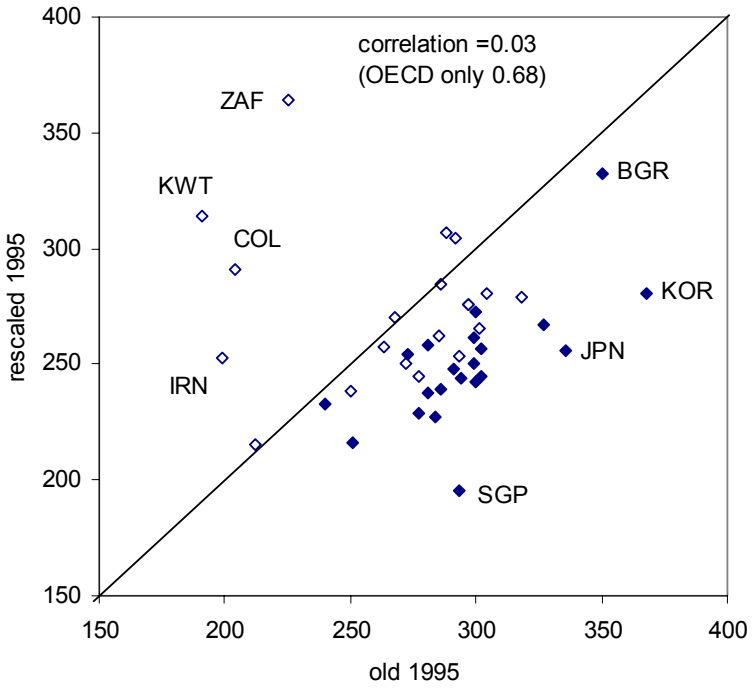
These graphs show that in both cases the correlation between old and new-scaled values is again very high, as for the median. However, the pattern of change between old and new scale results is not the same for both these quantiles. In the case of q5, countries with low values with the old scaling (which tends to be the non-OECD countries, although there are notable exceptions) get pushed even lower with the new scaling. Several of them get changed quite a long way – South Africa, Kuwait and Colombia. We surmise that this might be due to the new scaling method's extra parameter to account for guessing, i.e. allowing for guessing as one explanation of correct answers allows really poor ability to be revealed. On the other hand, countries in the top half of the distribution (the closed diamonds) get pushed up a bit, although the change is very small in this case.

In the case of q95, countries with the highest values on the old scale figures see the greatest reduction in the switch to the new scale values – Japan, Korea and Singapore. The countries with the lowest old scale values – South Africa, Kuwait and Colombia – see a slight rise on the new scale figures. Note that this is the opposite of the pattern for q5.

In general, the open-diamond countries have more change on q5 than on q95, and vice versa for the closed-diamond countries. The open-diamond countries tend to get pushed down on q5 but have only a modest reduction on q95 (and some of them are even pushed up). This means that, overall, the dispersion in maths scores, as measured by the *difference* between q95 and q5 goes up with the re-scaling for quite a few of these countries. The closed-diamond countries, on the other hand, tend to get pushed down on q95 but have very little change on q5 (a very small increase on average). This means that in their case the dispersion as measured by the difference between q95 and q5 goes down. So some countries have dispersion that goes up and some have dispersion that goes down.

The net result in terms of change in dispersion, measured by q95-q5, is given in **Figure 4**. The correlation coefficient is effectively zero. That is, there is no relationship between dispersion (measured in this way) in the 1995 old scale scores and the 1995 new scale scores. (Remember that the two sets of scores are derived from the same data.)

Figure 4: Old v. rescaled Q95-Q5, TIMSS 1995 8th-grade maths



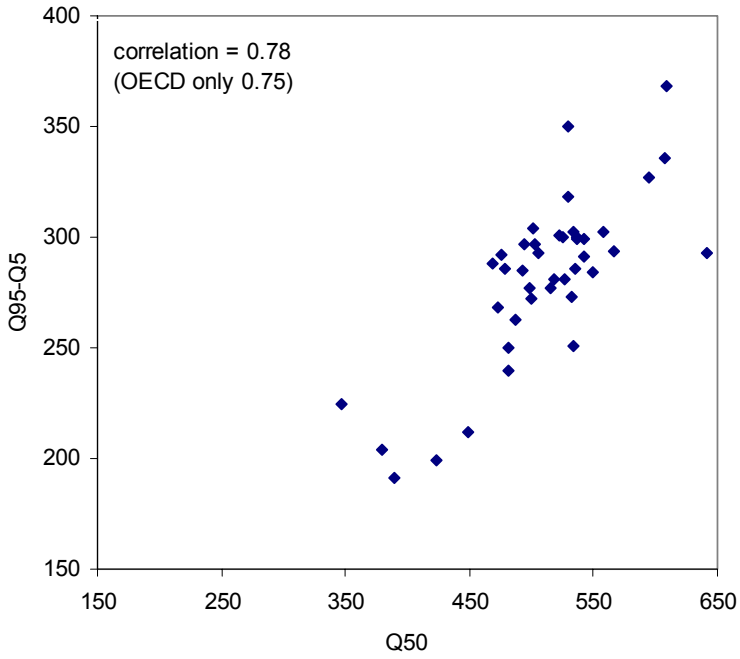
The open diamonds in Figure 4 are again the countries in the lower half of the distribution of q5 old-scale data in Figure 2. All the closed-diamond countries see a reduction in dispersion (they all lie below the 45 degree line). About half of the open-diamond countries see an increase and about half see a reduction. The change in the position of South Africa is dramatic.

The country with one of the smallest recorded values for dispersion with old scale scores, becomes the country with the largest dispersion – which seems more reasonable – with the new scale scores. The changes for Kuwait and Colombia are almost as striking. We then looked at how the re-scaling would influence one's view of whether dispersion in achievement rises or falls with average achievement (measured by the median), staying with the maths scores. This seems a fundamental issue for one's view of educational progress. Is there a trade-off between high average achievement and low inequality in achievement or do the two go hand in hand?

Figures 5 and **6** plot q95-q5 against the medians analysed in Figure 1, for both the old scale data (Figure 5) and new scale data (Figure 6). With the old scale data the conclusion is that countries with higher average achievement have *higher* dispersion in achievement. With the new scale data the opposite conclusion would be drawn – as average achievement rises, inequality in achievement *falls*, although it is worth

noting that the outliers of South Africa, Kuwait and Colombia on the one hand and Singapore on the other have a considerable impact on this conclusion.[7]

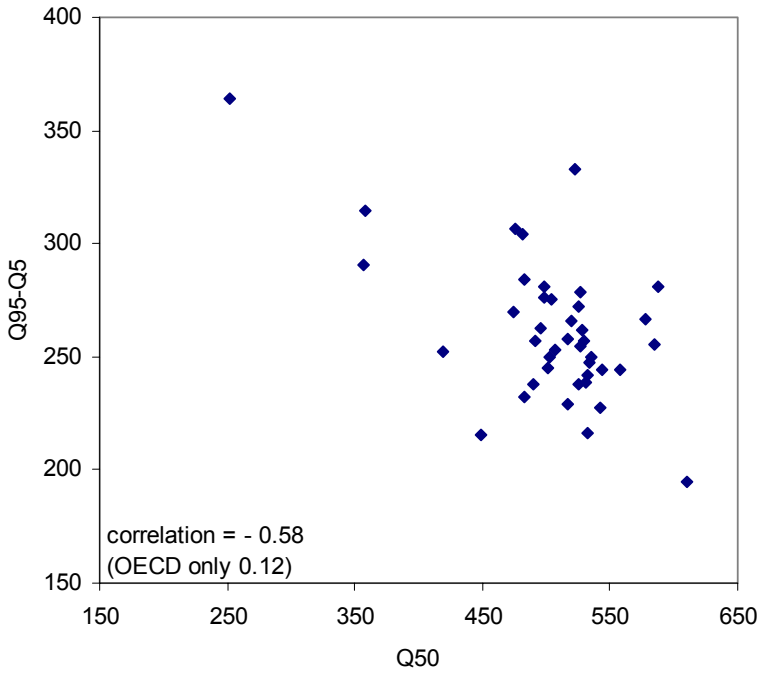
Figure 5: Q95-Q5 v. Q50, old TIMSS 1995 8th-grade maths



The sensitivity of the 'league table' for dispersion is much less for the OECD countries. In each of Figures 1 to 6, we have indicated the correlation coefficients for the OECD countries alone, which in general are higher achieving countries. The correlation between old and new scale values of q95-q5 is 0.68 for the OECD countries, compared to 0.03 for all countries (Figure 4). However, even for the OECD group, the conclusion on dispersion versus central tendency seems sensitive to the switch from old to new scale. The correlation between q95-q5 and q50 is 0.75 for the old scale scores (Figure 5) but only 0.12 for the new scale scores (Figure 6).

[7] Not surprisingly, the 1999 data give the same relationship as the 1995 new scale data, i.e. dispersion falls as average achievement rises.

Figure 6: Q95-Q5 v. Q50, rescaled TIMSS 1995 8th-grade maths



We repeated the whole exercise for the TIMSS science scores. A natural question is whether the results obtained for maths are peculiar to that subject. Again, the new and old scale quantiles themselves (q50, q5 and q95) are very highly correlated. The differences between q95 and q5, i.e. the dispersion in scores, are once more much less closely correlated, although in this case they are far from being unassociated – there is a correlation coefficient of 0.68 between the old and new scale values of q95-q5 for all countries in contrast to the figure of 0.03 for maths shown in Figure 4.

What can we conclude from this exercise? First, and most obviously, while the ranking of countries in TIMSS by the between-country differences (as summarised by the medians) is robust to the change in scaling method between 1995 and 1999, the ranking by the within-country differences (measured by q95-q5) is not in the case of maths, this being especially true if the full group of participating countries is considered rather than the OECD members alone. Second, and following on from the first conclusion, one has to 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.

b) *The calculation of standard errors*

Like any sample survey, the results from the achievement surveys are subject to the random variation generated by the sampling process. Moreover, the surveys have a complex sampling design so that the calculation of standard errors based on an assumption of pure random sampling will produce estimates of these standard errors that are biased downwards. As we have noted, TIMSS and PISA first sample schools and then randomly select whole classes within schools (TIMSS) or randomly sample all students within schools (PISA). Either way there is strong clustering of the data that if ignored will lead to the bias just described.

All this is well known and the survey reports are careful to both emphasise the problem and to provide estimates of standard errors for published summary statistics that allow for it, typically using a jackknife or jackknife-like procedure.[8] In the case of TIMSS, the supplied SPSS and SAS programmes referred to in Section 2 enable the user to implement the recommended jackknife procedures for the estimation of standard errors, whether for standard summary statistics such as the mean or for OLS regression coefficients. However, the use of the jackknife in SPSS is computationally burdensome in the case of regression with more than a handful of variables.

Users may also want to estimate other types of statistical model for which no programme is supplied. In the case of PISA there is no programme supplied to help with the estimation of standard errors but the advice on how to use the 'balanced repeated replication method', which is similar to the jackknife, can be followed for estimates of the standard errors for any procedure, although we suspect that not all users will carry this through.

We think it likely that many users will be using STATA rather than SPSS or SAS. As a result they will be drawn to use the STATA 'svy' ('survey') commands that allow for estimation of standard errors in the presence of (user declared) stratification and clustering where intra-cluster correlation can take a general (and unspecified) form. The 'svy' commands in STATA are now very extensive with versions available for all manner of linear and non-linear models, as well as for standard descriptive statistics.

[8] It should be noted that the published TIMSS 1999 maths and science reports contain errors in the standard errors of summary percentiles (q5, q25, q50, q75 and q95), resulting in them being substantially smaller than they should be. These errors are corrected in the .pdf versions available on the TIMSS website.

The danger here is that the user may believe that the 'svy' commands solve his or her problem completely. **Table 1** compares estimates of standard errors of the mean PISA reading score for a selection of countries, comparing (a) the published standard errors based on the balanced repeated replication (i.e. jackknife-like) method with (b) the standard errors obtained with STATA while assuming pure random sampling (i.e. ignoring the complex survey design), (c) those estimated with the 'svymean' command (declaring clusters but not strata), and (d) estimates obtained using the bootstrap, also available in STATA and another attractive feature of this computer package (we bootstrap with the clusters).

Table 1: Estimates of standard errors of the mean, PISA reading

	Mean	Standard Error of Mean			
	(published)	(a) published	(b) Simple random sample	(c) STATA 'svymean' command	(d) STATA Bootstrap 80 repl.
Australia	528	3.5	1.4	3.7	3.2
Germany	484	2.5	1.5	6.8	6.3
Italy	487	2.9	1.3	5.3	5.4
Japan	522	5.2	1.1	5.3	5.4
Portugal	470	4.5	1.4	5.4	5.4
United States	504	7.1	1.6	5.3	5.1

Note: Simple random sampling takes weights into account. The STATA 'svymean' command uses the average of the five plausible values and declares clustering. The bootstrap command uses 80 replications for the mean of the five plausible values and takes clustering into account.

The ratio of columns (a) to (b) provides an estimate of the design effect of PISA for the mean. The standard errors estimated by the survey organizers are two to three times larger in most cases than those produced under the assumption of pure random sampling, although in the case of Japan and the United States the ratio approaches 5. Use of the standard errors estimated under an assumption of pure random sampling could clearly lead very frequently to the wrong conclusions being drawn from hypothesis tests involving the mean (depending on the nature of the hypothesis).

How well does STATA's 'svymean' procedure do, by the benchmark of the published standard errors? In the case of the US, the estimate in column (c) lies in between those in columns (a) and (b). The conclusion one would reach is that the use of the 'svymean' procedure goes a long way to solving the problem but is still 'too small'. For all other countries, it appears to do the trick – Australia and Japan – or to lead to overshoot with the estimate in column (c) exceeding the published figures in column (a), which is the case in Germany, Italy and Portugal. We are unable to say why this is the case but it is possible that it is because we are not using the information on the survey strata when using the 'svymean' procedure.

This pattern is repeated for the other countries not in the table – typically the results of the 'svymean' procedure gives estimates that are a fair bit larger than those published but in some cases the increase is only small. On this evidence the 'svymean' procedure would lead one to err on the side of caution. The same applies to the use of the bootstrap since the estimates in column (d) are close to those in column (c).

Table 2 shows the results of a similar exercise with TIMSS 1999 data. Here we compare the results of an OLS regression for maths scores of grade 8 children with standard errors estimated (a) assuming pure random sampling i.e. with no allowance for clustering, (b) with the STATA 'svyreg' command that allows for clustering and (c) with the SPSS programme supplied to users by the survey organizers that incorporates the jackknife. We take a different selection of countries to that in Table 1. The explanatory variables included are dummy variables for gender (equal one if boy), mother's and father's education (equal one if at least secondary), and the number of books estimated by the child to be present in the home (equal to one if more than 100).

Table 2: Estimates of standard errors of regression coefficients, TIMSS 1999 maths

	Canada	Korea	New Zealand	Finland	Czech Republic
Gender	**5.8**	**7.9**	**- 4.7**	**9.7**	**15.7**
Simple random STATA	(2.8)**	(1.9)***	(3.5)	(3.9)**	(3.6)***
Svy STATA	(2.5)**	(2.6)***	(8.0)	(4.7)**	(4.3)***
Jackknife SPSS	(3.2)*	(3.1)***	(8.4)	(5.4)*	(5.3)***
Mother's edu	**17.1**	**10.5**	**12.6**	**20.2**	**13.8**
Simple random STATA	(5.6)***	(2.7)***	(4.9)**	(5.6)***	(4.5)***
Svy STATA	(3.9)***	(2.8)***	(5.2)**	(6.0)***	(4.4)***
Jackknife SPSS	(8.7)*	(3.4)***	(5.7)**	(8.1)***	(6.7)**
Father's edu	**12.6**	**11.1**	**10.6**	**15.9**	**17.1**
Simple random STATA	(4.7)***	(3.1)***	(4.4)**	(5.0)***	(4.5)***
Svy STATA	(4.9)**	(3.3)***	(4.8)**	(5.2)***	(4.2)***
Jackknife SPSS	(4.5)***	(4.1)***	(5.2)*	(7.6)**	(6.7)***
Books in HH	**8.0**	**21.1**	**23.3**	**11.2**	**17.0**
Simple random STATA	(1.3)***	(0.9)***	(1.5)***	(2.1)***	(2.0)***
Svy STATA	(1.4)***	(0.8)***	(2.0)***	(2.4)***	(2.3)***
Jackknife SPSS	(1.9)***	(1.2)***	(2.2)***	(2.7)***	(2.8)***
Constant	**481.3**	**501.9**	**399.0**	**464.9**	**436.8**
Simple random STATA	(6.9)***	(3.4)***	(6.2)***	(7.4)***	(7.7)***
Svy STATA	(12.2)***	(4.4)***	(9.0)***	(7.2)***	(8.5)***
Jackknife SPSS	(6.7)***	(5.4)***	(10.2)***	(8.8)***	(10.1)***
Observations	5422	5120	2156	1012	2645
R-squared	0.04	0.15	0.13	0.14	0.10

Note: Parameter estimates are given in bold and standard errors estimated with different methods in parenthesis.

* significant at 10%
** significant at 5%
*** significant at 1%

Simple random sample estimation of standard error takes weights into account. Simple random and svy calculations use the mean of the 5 plausible values. Dependent variable: math achievement; Independent variables: gender (0=girl, 1=boy), mother's edu, father's edu (0= not finished secondary, 1= finished secondary), books (1=0-10 books in household, 2=11-25 books, 3=26-100 books, 4=101-200 books, 5= more than 200 books)

In most cases the results are fairly clear: the jackknifed standard errors tend to be about 20-50% larger than those estimated under an assumption of pure random sampling, with those estimated with the STATA svyreg procedure coming somewhere in between. The exception is Canada where the svyreg procedure seems to lead to substantial overshoot, relative to those estimated with the jackknife. With this exception, the broad conclusion from Table 2 is that some caution is needed when using standard errors based on the svyreg procedure. For example, one would reject the null hypothesis of no gender differences in Finland at the 5% level when using the standard errors from the svyreg procedure, but at only the 10% level when using the jackknife estimates from the SPSS programme supplied by the survey organizers.

c) *The standard error of a difference in quantiles and the Bonferroni adjustment*

Lastly in this section we deal briefly with two other matters involving calculation or use of standard errors that we have encountered in our work.

First, we consider the standard error for a difference in quantiles, e.g. q95-q5. The TIMSS, PISA and IALS reports publish standard errors of selected quantiles. The issue rises as to how to calculate the standard error of a *difference* in these quantiles, one measure of the amount of dispersion in a distribution. Despite at least one measure of a difference in quantiles, the inter-quartile range (q75-q25), being a common textbook measure of dispersion, strangely enough the full procedure for deriving the standard error of the difference between any two quantiles seems not to be widely known.

The variance of this difference is obviously equal to the sum of the variances of each of the two quantiles of interest minus twice their covariance. The key issue here is the covariance (since estimates of the variances are known from the published estimated standard errors of the two quantiles). One possibility is to set this to zero. In this case the standard error of the difference in quantiles will be overestimated (assuming the covariance is positive), which is clearly a mistake on the right side of caution. However, it turns out that the covariance is easily recoverable from the standard errors of the quantiles.

The variance-covariance matrix for any set of quantiles is given in Kendall and Stuart (1969).[9] Let F be the cumulative density function of the distribution in question and f the probability density function. Let p_i, i=1..k be proportions, where $0<p_1 < p_2 < p_k < 1$. Given our particular interest, we assume that k=19 and $p_1 = 0.05$ and $p_{19} = 0.95$. Kendall and Stuart give the variance-covariance matrix of quantiles from a random sample of size n as:

[9] See also Beach and Davidson (1983).

$$\begin{array}{ccc} \frac{p_1.(1-p_1)}{n.(f_1)^2} & \cdots & \frac{p_1(1-p_k)}{n.f_1.f_k} \\ \vdots & & \vdots \\ \frac{p_1.(1-pk)}{n.f_1f_k} & \cdots & \frac{p_k(1-p_k)}{n.(f_k)^2} \end{array}$$

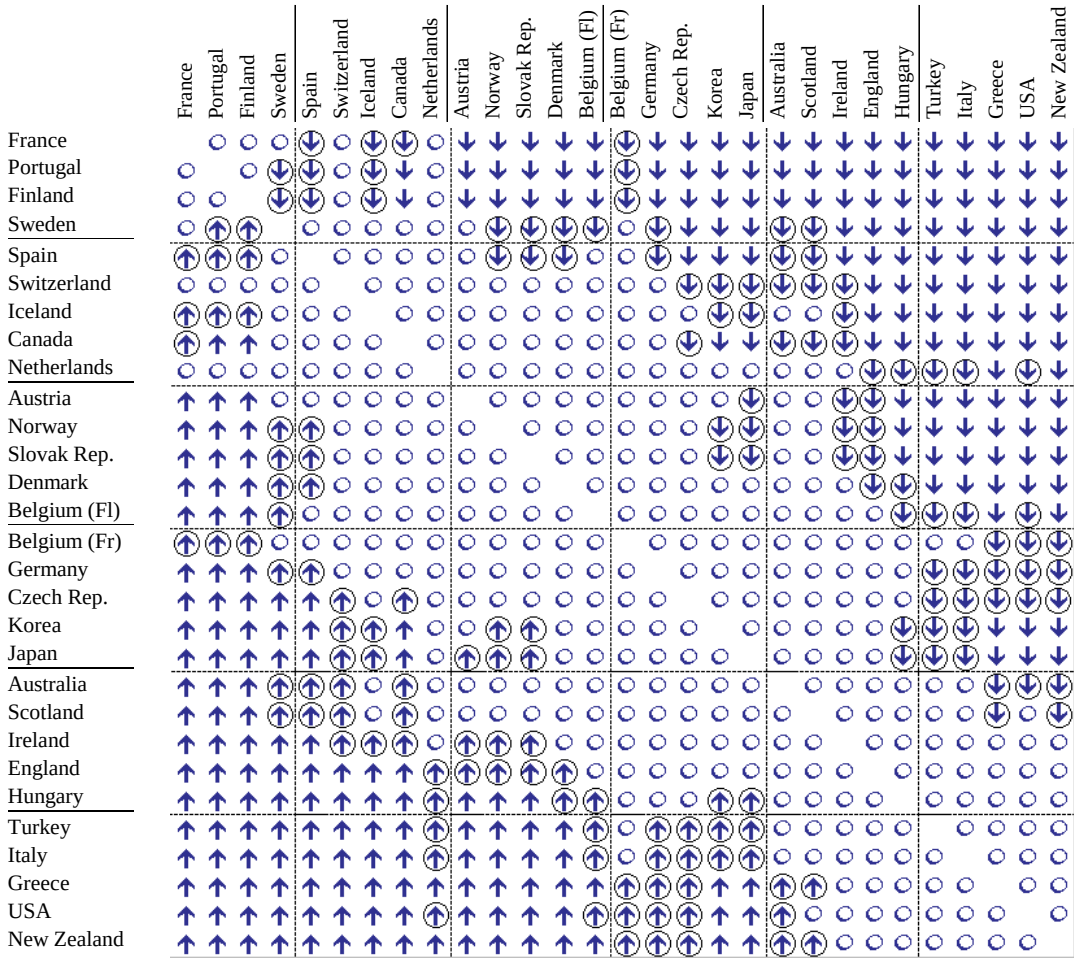
The unknown information in the off-diagonal elements, the covariances, are the f_i. But since estimates are published of the square roots of the diagonal elements, the variances (i.e. the standard errors), the f_i can be recovered from these, e.g. $f_1 = [p_1.(1-p_1) / n.var_1]^{0.5}$. (This expression for f_i contains n, the sample size, and this must also be considered as unknown since in principle it is the effective rather than the actual sample size, given the sample design. However, n falls out of the expression for the denominator of the off-diagonal elements, e.g. $n.f_1.f_k$, leaving an expression involving only known elements, i.e. the p's and the var's)

Second, we comment on the Bonferroni adjustment for multiple comparisons typically made in the survey reports. Imagine we want to test for significant differences between countries in the amount of dispersion of test scores, as measured by the standard deviation. We set a significance level of 5%. This is equivalent to specifying a probability of Type 1 error that we are prepared to accept, that is of the probability of erroneously rejecting the null hypothesis. (In other words, we accept a 5% chance that we will conclude an observed difference in dispersion between the two countries is significant when it is really due to sampling variation.)

This possibility of Type 1 error will occur *each time* we make a pairwise comparison of Country A with another country, Country B, C, D, E etc. This makes the cumulative probability of making a Type 1 error when testing for differences between Country A and any other country quite high. (And adding more countries to the sample would have the effect therefore of increasing the probability of making a Type 1 error at least once.) We may instead want to restrict to 5% the chance that we *ever* make a Type 1 error when comparing Country A with any other country. This involves making the so-called Bonferroni adjustment to the tests, which in practice implies setting a critical p-value in testing at the 5% level of 0.05/N where N is the number of comparisons (number of countries minus one).

Table 3 shows all pairwise tests of differences in standard deviations of maths scores for OECD countries in TIMSS, pooling 1995 (new scale) and 1999 8th grade results and taking the later year for those countries in both rounds of the survey. The arrows that are circled in the diagram are those where the difference between two countries is <u>not</u> statistically significant when we make the Bonferroni adjustment.

Table 3: Significance tests, TIMSS maths standard deviation



Notes:

- ❍ no statistically significant difference at the 5% level
- 🡻 value of country in row significantly smaller than in country in column, with Bonferroni adjustment for 28 multiple comparisons
- ⓓ value significantly smaller than in comparison country without Bonferroni adjustment, but not significantly different with
- 🡹 value (of country in row) significantly larger than in comparison country (column)
- ⓤ value significantly larger than in comparison country without Bonferroni adjustment, but not significantly different with

The Bonferroni adjustment is applied in the published TIMSS reports when testing for differences in the mean and in the proportion of the sample reaching a given fixed level of achievement such as the international median. (The published reports do not give tests of differences in dispersion.) But there are arguments both for and against its use. The argument in favour has just been stated. The argument against can be seen when the Bonferroni adjustment is described in a different way. What it implies is a null hypothesis that dispersion in Country A is the same as in Country B and in Country C and in Country D etc. Just one arrow anywhere along a row (an uncircled arrow) is sufficient to reject this null. Table 3 shows that the null is rejected for all but one country. But this is a very cautious approach to testing the country differences. Having rejected the null just described for Country A we naturally want to then go to look at the pairwise differences. For this purpose the tests without the Bonferroni adjustment seem appropriate and in our view they are the ones to focus on.

The table also shows how important it is to test for significance in the observed differences rather than accepting them at face value. Without the Bonferroni adjustment for multiple comparisons, half (54%) of the pairwise differences for maths are significant at the 5% level while with the adjustment the figure drops to only a third (34%). The zones without arrows in Table 3 mean, for example, that we cannot reject the hypothesis that all the eight most unequal countries have the same degree of dispersion in maths achievement. The standard deviation for New Zealand of 89 is not significantly different from that for Ireland of 83.

On the other hand, clear water does come between those groups of countries at either ends of the ranking. Dispersion in maths scores in all eight countries from New Zealand to Ireland is significantly higher than that in the eight least unequal countries, Canada to France (not applying the Bonferroni adjustment). Were it not for the Ireland-Netherlands comparison, where the null cannot be rejected, one could add the four countries Slovak Republic to the Netherlands to this latter list.

4. Methods for presenting summary results

Achievement scores have no natural metric. They are numbers on a scale chosen by the survey organizers, with the typical choice that the mean among all participants is equal to 500 and the standard deviation to 100. Conveying their meaning is therefore not a straightforward task.

The TIMSS, PISA and IALS survey reports all present their results starting with two main figures: (a) the percentage of persons (students or adults) reaching given international benchmarks or performing at specific proficiency/ literacy levels in each country[10]; and (b) the distribution of achievement or performance/literacy scores in each country, where countries are ranked by mean scores[11]. We comment on each of these before turning to methods that we have found useful in our own work.

a) *Benchmarks and levels – 'absolute' educational disadvantage*

The percentage of students in each country whose performance falls below a fixed benchmark score can be interpreted as a measure of 'absolute' educational disadvantage. The idea here is that the chosen benchmark level represents the same level of achievement in each country. Countries in which a large proportion of students fall below a given level of competence clearly have a cause for concern over future productivity and competitiveness. (Of course, the fewer people that reach the benchmark level of achievement in any country the more valuable to the individual is likely to be the achieving of that level, in terms of access to higher education or to wages in the labour market. That, however, is a different issue.)

In **Figure 7**, the dark bars show the three countries at the top and the three at the bottom of the PISA league table of OECD countries on such a measure of absolute advantage.[12] The numbers refer to the percentage of 15-year-olds in each country who fall below PISA's Level 2 for reading literacy. Such students, according to the PISA organizers, are "unable to solve basic reading tasks, such as locating straightforward information, making low-level inferences of various types, working out what a well-defined part of a text means, and using some outside knowledge to understand it." And as the figure shows, the percentage of students judged to be disadvantaged in this way varies considerably – from 6% or 7% in Korea and Finland to more than 20% in Hungary, Greece and Portugal.

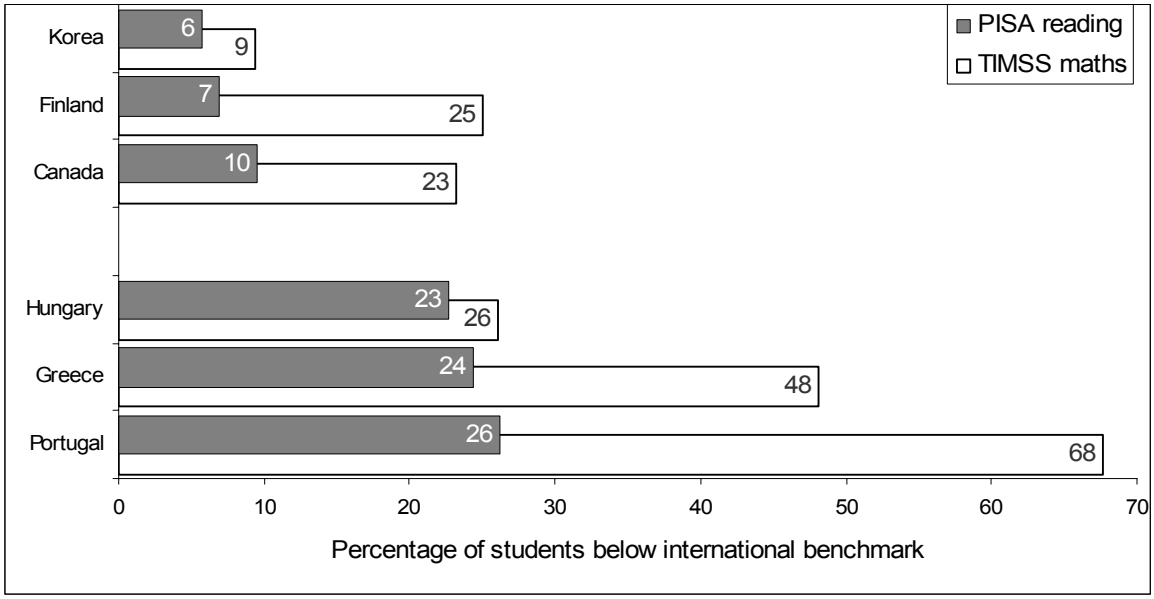
---

[10] Exhibit 1.6 in TIMSS 1999 reports, Figure 2.3 in PISA report, Figure 2.2 in IALS.

[11] Exhibit 1.1 in TIMSS 1999 reports, Figure 2.5 in PISA report, Figure 2.1 in IALS.

[12] We exclude Luxembourg, Mexico and Poland for this purpose since we wish to compare the figures with TIMSS, which did not include these three countries.

Figure 7: Benchmarks and levels



The light bars in Figure 7 show a measure of absolute disadvantage based on TIMSS: the percentage of 8th grade students in each country who, according to the TIMSS organizers, are "unable to apply basic mathematical knowledge in straightforward situations". These percentages are in general higher than for the PISA benchmark (and are so in every case for the six countries in the table), although they cannot be compared directly with the equivalent PISA ones. But in general, we see that countries which do relatively well or badly in PISA tend to do so also in TIMSS.

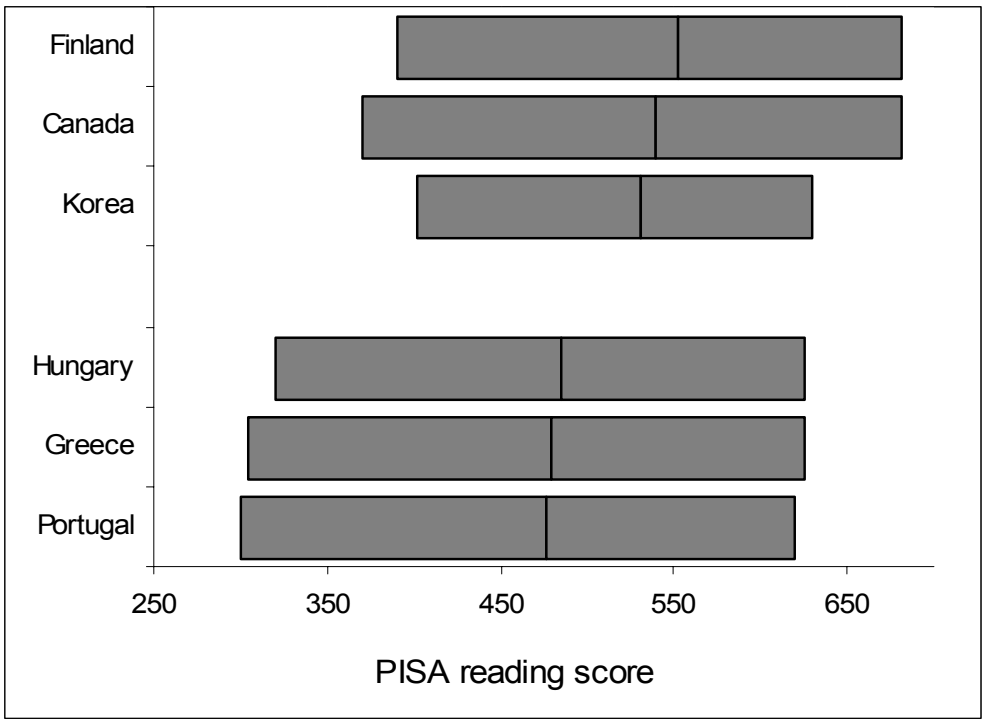
These measures of absolute under-achievement have an obvious attraction but one needs to realise that the user is taking the benchmark levels 'on trust'. Consider the analogy with an international poverty line fixed as a given level of U.S. dollars or Euros per day in purchasing power terms, as used for example by the World Bank. A figure of say two dollars per day may be an arbitrary choice but at least everyone has a feel for what two dollars means in terms of what can be bought with this sum in their own country. On the other hand, the PISA and TIMSS benchmarks just described have the attraction of not seeming arbitrary – they pertain to what seem 'real' threshold levels of achievement. But it is the survey organizers who are the judges of what scores should be the threshold levels and one has to trust that their judgement is correct. In the case of the PISA benchmark described above, the threshold level is defined *a priori* but in the case of TIMSS the benchmark simply corresponds to the median among all students in all participating countries. This risks appearing as arbitrary as the choice of a poverty line as two rather than say three dollars a day.

PISA, TIMSS and IALS all have more than one of these benchmark levels – there are typically four or five in each survey corresponding to different scores judged by the survey organizers to correspond to different threshold levels of achievement. In effect, this provides what might be called a 'partial' metric. It is rather like having a tape measure to measure people's height which is blank but for a few unevenly spaced marks. 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 height of two people.

b) *Distribution of achievement*

Figure 8 illustrates the other principal method used by the TIMSS, PISA and IALS survey organizers to present basic results. It shows the dispersion in PISA reading literacy scores within the same six countries included in Figure 7. The bars extend from the 5th to the 95th percentiles of the national distributions, with the lines approximately at the middle of each bar corresponding to the 50th percentile, or median. (We do not include on the diagram an indication of the 25th and 75th percentiles that are also typically included in the survey report versions of this diagram.)

Figure 8: Distribution of scores, PISA reading



In such a diagram we can compare the differences between countries according to their median scores with the differences within countries. For example, the difference between the medians of high-scoring Finland and low-scoring Portugal is about 75 score points, while the average difference within countries is almost 300 points, which is about four times as much. This helps put the between-country differences in perspective. The degree of inequality in educational outcomes within a country is an important summary statistic since most governments are concerned about education as a means of furthering equality of opportunity and social cohesion.

The ranking in this figure is quite similar to that relating to benchmarks and levels (absolute disadvantage), since the percentage below a benchmark level is highly correlated with the median (and the mean). For example, the correlation of the percentage at or below level 1 and the median in PISA reading literacy for all OECD countries is 0.96. But a ranking by the lengths of the boxes – the degree of inequality within each country – would show a different story.

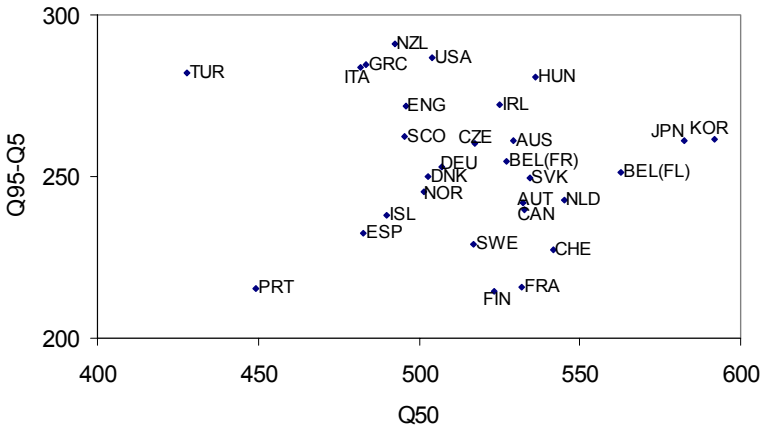
One problem with this kind of diagram is that it is difficult to compare directly individual country values of dispersion in achievement within countries with that between them, that is, to compare the lengths of the boxes (the difference between 95th and 5th percentile) with the values of the median. Also, we get no idea of the shapes of the distributions, and the temptation may be to see them as uniform, with people evenly spread between the extreme percentiles.

c) *Scattergrams of dispersion versus central tendency*

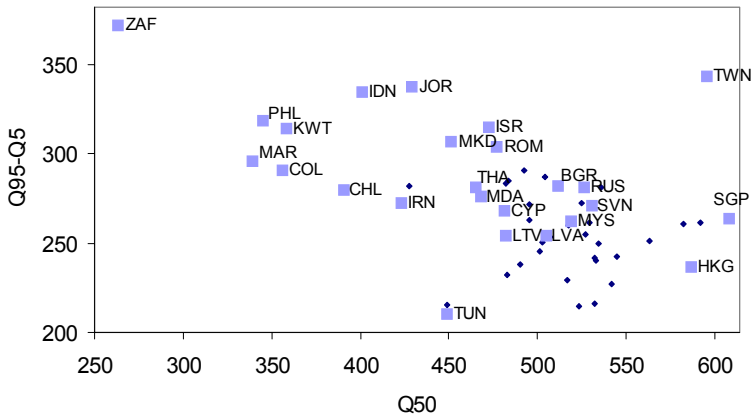
We now turn to methods of presentation we have found useful when presenting basic results in order to augment those used by the survey organizers. **Figure 9** shows a scattergram of dispersion versus central tendency in TIMSS maths, showing each country's value of the difference between 95th and 5th percentiles and its median.

Figure 9: Q95-Q5 v. Q50, TIMSS maths

OECD countries



All countries

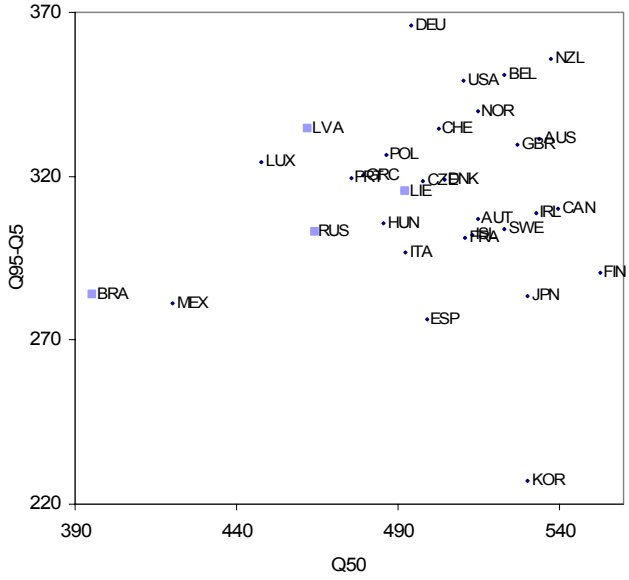


Small diamonds = OECD countries
Squares = other countries

Figure 10 does the same for PISA reading scores. We have already shown such a diagram (for 1995 data only) in Figure 5 and 6 when discussing methods of 'scaling' the raw data. As there, the idea behind the diagram is to see more clearly the relationship between dispersion and central tendency. Unlike Figure 8, which also contains information on both measures, the scattergram in Figure 9 gives them equal weight – there is no ranking on one of them as in Figure 8. (Of course, the earlier investigation of scaling might make one want to give less weight in presentation to dispersion, on the grounds that dispersion seems more sensitive to

scaling method, but that in principle is a different issue.) It is much easier in Figures 9 and 10 to spot the outliers in terms of dispersion than in Figure 8 and (conditional on one being happy about the scaling method) one can use Figures 9 and 10 to draw simple conclusions on whether there seems to be a trade-off between inequality and central tendency or not.

Figure 10: Q95-Q5 v. Q50, PISA reading



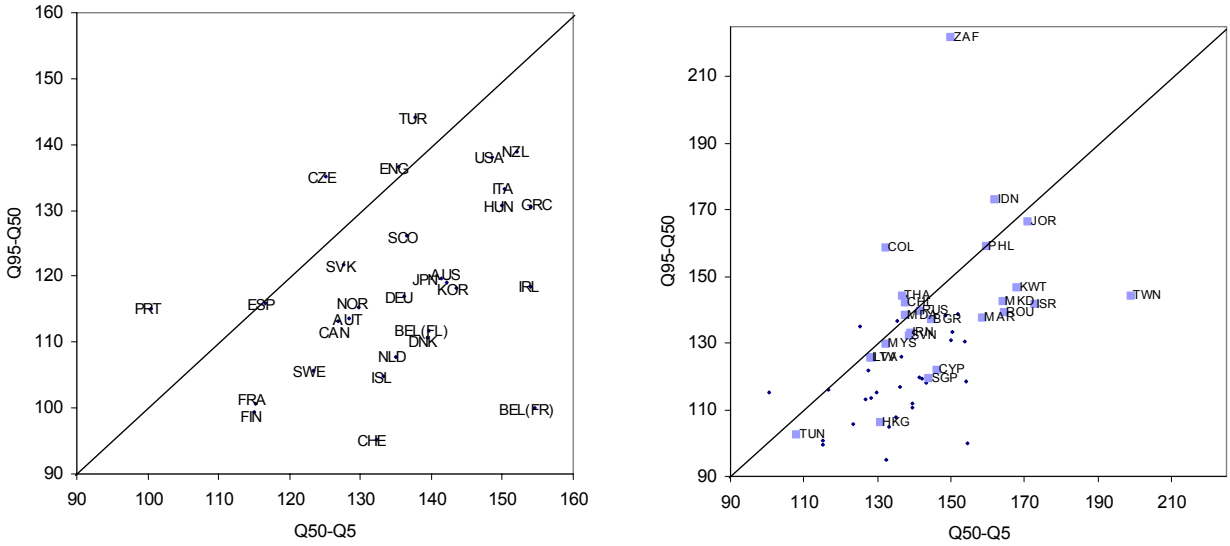
d) *'Relative' educational disadvantage*

We have noted above that the degree of inequality in achievement within a country is of obvious interest to any discussion of social cohesion and educational opportunity. However, views about the undesirability of inequality may depend on which end of the distribution one is considering.
One might argue that inequality at the top has some desirable features if it reflects the fact that high innate achievers are in an educational system that allows them to demonstrate their ability (although there are counter arguments that we don't go into here). However, most people would certainly agree that higher inequality at the bottom of the distribution is a bad thing.

We can think of this as implying higher 'relative' educational disadvantage in a country. In our work therefore we have taken the difference between the median and the 5th percentile as a measure of this relative disadvantage. The greater this difference, the further away from the *national* norm are a country's low achievers. Where this difference is greatest, the weakest students are being allowed to fall further behind the average than in other countries.

Figure 11 plots this measure of relative disadvantage (q50-q5) for TIMSS maths scores on the horizontal axis against the difference between the median and the 95th percentile (q95-q50) on the vertical axis. The former exceeds the latter in most countries: in contrast with income and wage distributions, for example, the achievement data are negatively skewed with somewhat more inequality at the bottom than at the top. Portugal, Turkey and the Czech Republic are exceptions as are Spain and England where the 5th and 95th percentiles are equidistant from the median. This general pattern may reflect in part the design of the test and the scaling procedures. Of more interest is the degree of correlation between inequality in the two halves of the distribution.

Figure 11: Q50-Q5 v. Q95-Q50, TIMSS maths (OECD only and all countries)



This correlation turns out to be not that high. It is true that high inequality countries like the USA and New Zealand have high dispersion at both top and bottom of the distribution (relative to other countries) and low inequality countries like France and Finland have low dispersion in both halves (again, relative to other countries). But there are other examples where one gets a lot more insight from looking at the two halves separately. For example, Ireland's high overall inequality is shown to result from a long lower tail; the upper half of the Irish distribution is about the same length as that in Spain, a rather low inequality country overall when measured by q95-q5. The French-speaking part of Belgium has the longest lower tail of all, but one of the shortest upper tails.

While we have found the focus on q50-q5 easy to justify in principle, we should counsel that in practice one needs to be careful to take account of sampling error. Consider all pairwise comparisons of OECD countries' values of q50-q5 in TIMSS maths. Without the Bonferroni correction (see above) we reject the null hypothesis of equality of these differences in only a quarter of cases for maths. With the correction the null is rejected in only 1 in 10 tests. Nevertheless, some reasonably firm conclusions can be made about the end of the distribution. Without the Bonferroni adjustment, the great majority of tests between the seven countries with the largest values of maths q50-q5, French Belgium to USA, and the 10 countries with the smallest values, Norway to Portugal, lead to rejection of the null.

What do these results imply about the robustness of the country *rankings* for q50-q5? We investigate this issue through Monte Carlo simulations for the OECD countries.[13] For each simulation, we compare the rank on q50-q5 of each country included in the bottom half of Figure 11 with its rank in a new set of country values of q50-q5. This new set is created as follows: for each country we construct a new value of q50-q5 equal to the observed value (the value in Figure 11) plus the value of a random draw from a normal distribution with a mean equal to the standard error of the actual value of q50-q5. After each simulation we calculate summary statistics describing the change in ranks across the 29 countries. We carry out this simulation 10,000 times. The average absolute change in rank (averaging across the 29 countries) is equal to 3.4 places.

e) *Graphical representations: Kernel density plots – and the search for a 'metric'*

We noted earlier that the form of presentation in Figure 8 does not reveal much about the shape of each country's distribution and the same is of course true about Figure 10 as well. One method of presentation that we have found very effective (including in seminar presentations) is simply to reveal the full shape of a country's distribution in two dimensions.
One possibility would be to do this through the use of histograms. But we have found a much more effective method is with Kernel density estimates.[14]

This is a particularly attractive – and revealing – form of presentation for TIMSS data. In the case of TIMSS, we can exploit the coverage by the survey of two consecutive school grades, 7th grade and 8th grade. We explained in Section 2 that TIMSS 1995 covered both these grades but so far in this paper we have referred only to the 8th-grade results. And we have not used at all the results for the 3rd and 4th grades which were also included in the 1995 survey.

---

13 This paragraph reports on results obtained by Robert Waldmann, University of Tor Vergata, Rome, which form part of joint unpublished work.

14 These are obtained very easily using STATA.

Figure 12 shows the distribution of maths scores in 1995 separately for 7th and 8th graders in Portugal (a country with low inequality of TIMSS achievement), while **Figure 13** shows the 3rd and 4th grade distributions in Canada. (The Figure 12 distributions are for the re-scaled 1995 data, i.e. using the same scaling procedure as that applied to the 1999 data.) These diagrams underline vividly the extent of the within-country distribution. Up and down the two distributions, the gap between 7th and 8th graders is about 30 points. In contrast, the difference between 5th and 95th percentiles for the older group is around 220 points. In other words, the within-country difference, measured by q95-q5, is about *seven times* the progression between years at most points of the distribution. In effect, this diagram suggests a complete rather than a partial metric – a measuring rod for any difference in scores that is a multiple or a fraction of the difference between grades, something that is quite easily understood.

Figure 12: Kernel density plots for 7th and 8th-grades

Distribution of TIMSS maths scores in Portugal for the two adjacent grades with most 13-year-olds

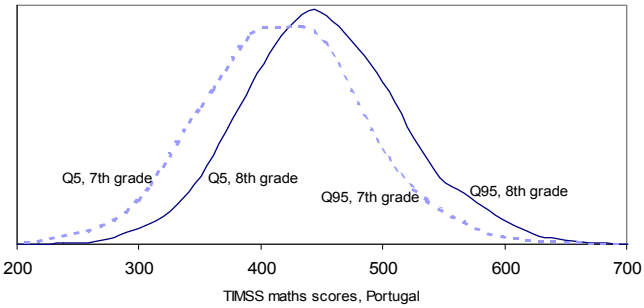
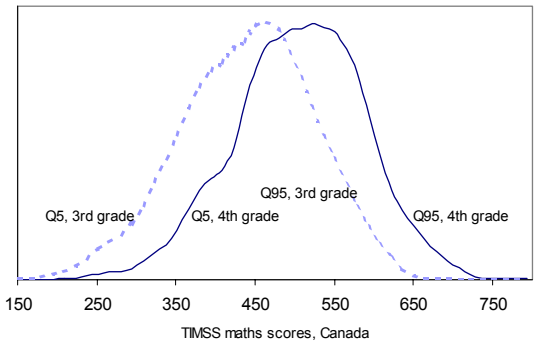


Figure 13: Kernel density plots for 3rd and 4th-grades

Distribution of TIMSS maths scores in Canada for the two adjacent grades with most 9-year-olds



An analogous calculation for the younger age group can be made using Figure 13, underlying again the extent of inequality even in the much younger age groups.

Of course, the extent of the difference in scores between grades differs from country to country. The difference between mean (re-scaled) scores between 7th and 8th graders varies from around 20 points in Korea, Germany, Ireland and the Netherlands (and as few as 7 points in the Flemish part of Belgium) to almost 40 points in France, Italy and Spain (and 46 points in Greece). And the difference is not exactly the same at all points in the distribution. So one is still some way from having a single complete metric, applicable in all countries. Nevertheless, these grade differences seem useful as a reasonable national metric for the TIMSS data and we have found that their use really does help people appreciate the extent of the within-country differences in achievement.

5. Comparing the results of the different surveys

Each study – TIMSS, PISA, IALS and others not covered in this paper – has adopted a different approach and emphasis. Each has been challenged on one ground or the other: is the testing culturally and linguistically neutral?[15] How is a "soft" value like literacy to be defined and measured? Are curriculum differences adequately taken into account? Is the sample representative of the target population? Are the students under test similarly motivated? Can performance in an achievement test reflect general ability in the full sense of the word (including the ability to apply knowledge learned)?

All of these problems may vary from survey to survey. Instead of basing all one's conclusions about what children learn on one individual survey taken in isolation, we think it important to compare the different surveys' results to see if a robust picture of the subject under investigation appears. Only in this way can full confidence in each survey's results be achieved. Of course, the different surveys aim to measure somewhat different things. Perhaps one should therefore not be surprised by any differences in the pattern of results from survey to survey. Nevertheless one should clearly wish to know whether there is agreement or not between the different surveys so as to avoid making conclusions that are heavily dependent on the use of one survey alone. In our view, there has been insufficient comparison to date of the different surveys' results. In some cases this is not possible since some subjects or background variables may only be covered by a particular survey. However, the general pattern of results in terms of basic results on levels of achievement/literacy can and should be compared.

We start by (a) discussing ways to *compare* basic results. We then (b) show one way to *combine* the results of the different surveys that we have found useful.

a) *Traffic lights*

One obvious and conventional way of comparing the results of different surveys is to look at correlation matrices for the basic results on central tendency and dispersion. This is the subject of **Table 4** where we use TIMSS and IALS (results for those aged 16-25 years only) for the 20 countries in both surveys. We show correlations for both the median (mean for IALS) and for the difference between 95th and 5th percentiles.

[15] Blum et al (2001) consider the experience of France's experience in IALS and among other things make critical comparison of the French language questionnaire used in France and that used in Switzerland. (France originally participated and then later withdrew.)

Table 4: Correlation matrices, TIMSS and IALS (aged 16-25)

	TIMSS (median)		IALS (mean)		
	Maths	Science	Quant.	Docum.	Prose
Maths	1				
Science	0.88	1			
Quantitative	0.72	0.57	1		
Document	0.62	0.48	0.93	1	
Prose	0.59	0.51	0.83	0.93	1

	TIMSS (Q95-Q5)		IALS (Q95-Q5)		
	Maths	Science	Quant.	Docum.	Prose
Maths	1				
Science	0.74	1			
Quantitative	0.71	0.62	1		
Document	0.51	0.54	0.87	1	
Prose	0.45	0.53	0.87	0.86	1

Looking first at central tendency, as one might expect, the TIMSS maths median for each country is highly correlated with the TIMSS science median (0.88). And it is also correlated well with the IALS quantitative scale mean (0.71), which is encouraging, while the correlation is less so for the IALS document and prose scales (0.62 and 0.59). Likewise, the TIMSS science median correlates better with the IALS quantitative score than with the document or prose scores.

Finally, the different IALS scales correlate more highly among each other than with the medians of either of the TIMSS subjects, repeating the pattern for the TIMSS scores.

The correlation matrix for dispersion shows the same broad pattern as that for central tendency, which is again encouraging for both surveys. Most of the correlations are somewhat lower in absolute size but this does not seem surprising (dispersion being harder to estimate well than central tendency). All but one exceed 0.5.

Table 5 shows an alternative method of comparing the results, something we have called a "traffic lights" diagram.[16] The table again refers to the 20 countries in TIMSS and IALS. The purpose is to show both the general pattern of correlation with colour – or as shown here in shades of black and white – and at the same time allow one to see how figures for particular countries compare across surveys. Moreover, the eye arguably picks up the general pattern of correlation among all the subjects and surveys better than it does in Table 4, which simply shows the pairwise correlations. The left hand side of the diagram shows the comparison for central tendency (median or mean) while the right hand side shows comparisons for dispersion (q95-q5).

Table 5: Traffic lights, TIMSS and IALS (aged 16-25)

	TIMSS (median)		IALS (mean)				TIMSS (Q95-Q5)		IALS (Q95-Q5)		
	Maths	Science	Quant.	Docum.	Prose		Maths	Science	Quant.	Docum.	Prose
Belgium (Fl)	563	539	300	300	288	Finland	214	255	130	141	130
Netherlands	545	551	294	300	291	Portugal	216	245	156	161	172
Switzerland	542	516	292	294	286	Switzerland	227	263	167	171	164
Hungary	536	556	282	267	262	Sweden	229	251	159	156	158
Canada	533	534	284	294	275	Canada	240	255	166	190	163
Slovenia	531	534	271	265	261	Netherlands	243	251	156	142	135
Australia	529	544	280	285	284	Norway	245	243	155	155	130
Ireland	525	523	274	272	277	Denmark	250	275	139	132	105
Finland	523	536	298	314	312	Belgium (Fl)	251	228	166	144	156
Czech Republic	517	539	303	295	278	Germany	253	300	139	138	139
Sweden	517	526	309	314	312	Czech Republic	260	262	148	162	127
Germany	507	527	297	294	283	Australia	261	284	167	157	159
USA	504	520	261	263	264	Slovenia	271	278	184	171	167
Denmark	503	478	301	306	283	UK	272	297	211	204	190
Norway	501	517	300	309	303	Ireland	272	295	190	177	170
UK	496	540	266	276	275	Chile	280	289	180	146	154
New Zealand	493	515	271	275	268	Hungary	281	275	184	184	147
Italy	482	496	269	268	270	Italy	284	287	170	161	163
Portugal	449	472	261	255	240	USA	287	319	230	233	225
Chile	391	423	229	237	241	New Zealand	291	304	201	210	211

Note: United Kingdom refers to England for TIMSS. Belgium (Fl) does not include the Flemish community in Brussels for IALS.

[16] We should acknowledge this as an idea for multidimensional comparison suggested by Bruce Bradbury, UNSW, Sydney, in earlier work together on child poverty rates in OECD countries.

In each case, countries are ordered by the values in the first column – TIMSS maths – and are divided into three groups and colour-coded on this basis. Dark denotes the worst performing countries, medium the average performers, and light the best (assuming lower means and higher variances as less desirable).

Reflecting the correlations in Table 4, we can see that in general the bottom third of countries in TIMSS maths are consistently dark-coloured according to all or most subjects and surveys – there is a band of dark colour running across the bottom of each side of the diagram. Three countries are always in the worst third in the case of central tendency and four in the case of dispersion. Some exceptions are easily identified. For example, Norway is in the worst third of values of central tendency for both TIMSS maths and science but in the best third in each of the three IALS subjects. On dispersion, Germany, Canada, Chile and Hungary are all found in each of the best, middle and worst thirds at least once.

b) *Average ranks*

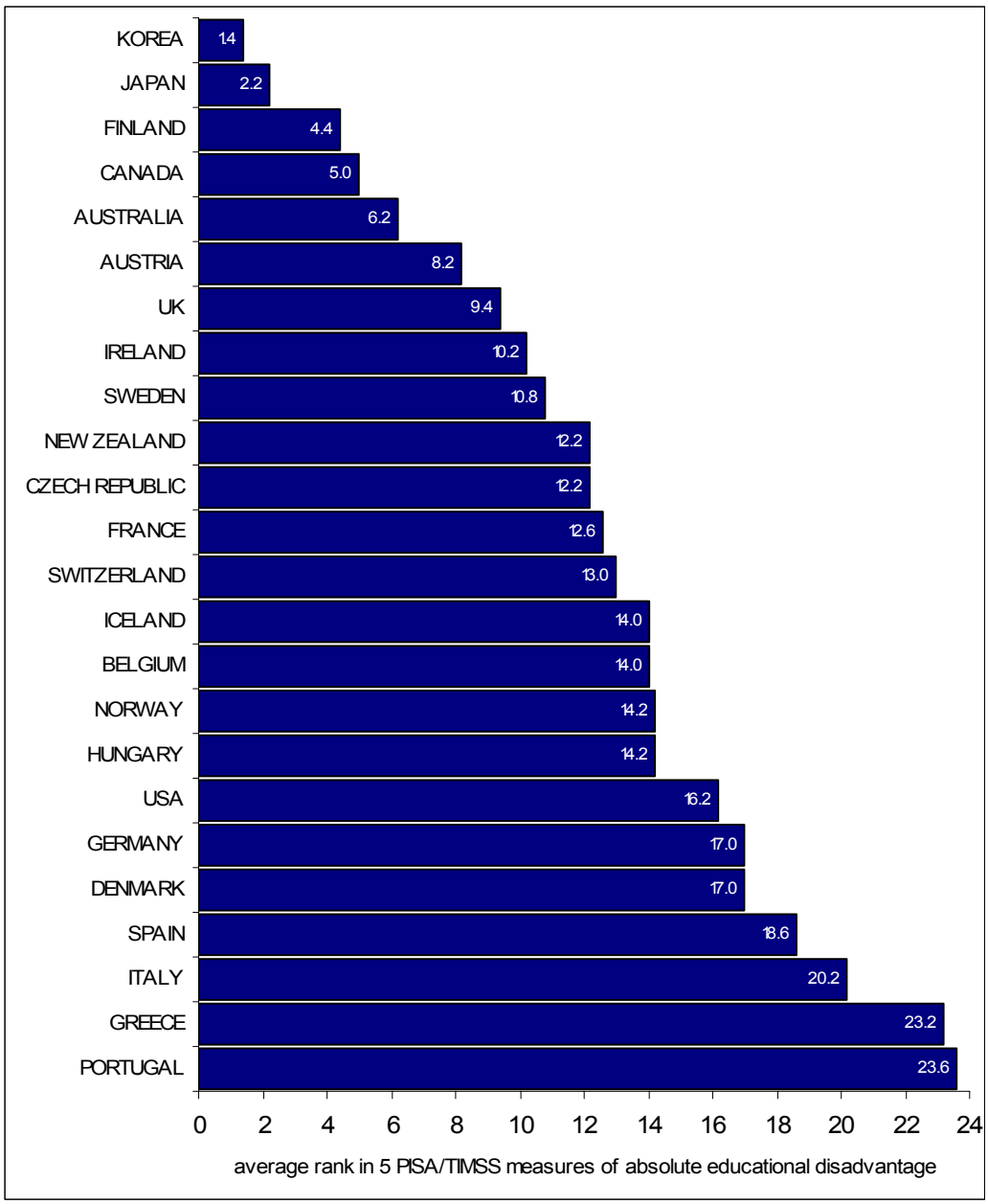
Summary statistics of achievement or literacy are not easily combined across the different surveys into one number. We have described how the 'scaling' process involves assigning the value 500 to the mean among all students in all countries and 100 to the standard deviation. On the face of it this should allow summary statistics to be easily combined so that one might, for example, consider the average mean score of each country across the different surveys. However, it needs to be borne in mind that the pool of participating countries varies from survey to survey and the logic is that this affects a country's results. Scoring 510 on average in a survey in which there are many countries with a weak level of achievement is more impressive than scoring 490 when the other countries are all strong achievers. This will restrict comparability. One possibility would be to calculate Z-scores for each measure for each country in the pool of countries participating in all the surveys in question and then to form an index of average Z-scores for each country. Here we choose instead another alternative based on ranks alone.

Figure 14 shows the *average rank* of each country in five different rankings of country performance in TIMSS and PISA. There are 24 countries entered in these (OECD only) rankings. That is, we look at ranks within this fixed group of 24, rather than among all countries participating in the surveys in question.

Performance is measured as the percentage of children scoring below a fixed international benchmark in reading literacy (below PISA literacy level 2), maths and science literacy (lower quartile of all children in OECD countries in PISA 2000), maths and science 8th-grade achievement (median of all children in all countries in TIMSS 1999, also for the countries with 1995 data). Children below these benchmarks are deemed to be at a disadvantage, as described above in Section 4.

If the different surveys produced wildly differing rankings then the averaging of the ranks would tend to produce a summary statistic with little variation. A low rank in one league table would likely be balanced by a high rank in another, so leaving all 24 countries clustered around an average rank of 12. Although a country in the middle of the table may arrive at that position either by being consistently around the middle of the individual league tables or by riding very high in one table and very low in another, being at the top or bottom ends of the table (with a very high or very low average rank) will be achieved only by scoring consistently well or consistently badly in the individual tables. We have seen from Tables 4 and 5 that there is reasonable agreement between TIMSS and IALS. On the evidence of Figure 14, this seems also to be the case for TIMSS and PISA since the *average* ranks show large variation. Japan and Korea for example do consistently very well, ranking on average 2.2 and 1.4 on the five different league tables, while Greece and Portugal consistently very badly, ranking on average at 23.2 and 23.6 out of 24.

Figure 14: Average ranks, TIMSS and PISA



6. Conclusions

The rest of this decade will see continued development of new international surveys of learning achievement and functional literacy. Users will have more and more data available to them, both in terms of summary statistics and analyses in published reports and in terms of microdata sets available for secondary analysis. In this situation it is important that users are able to draw on each others' experiences of using the surveys as well as on the formal user guides produced by the survey organizers. This paper is intended to contribute to such a process.

A range of issues have been dealt with in the paper. The most important of these are: (i) the robustness of results to the method used for 'scaling' the data (aggregating the answers for each individual into a single score); (ii) methods for presenting summary statistics; and (iii) comparison between the surveys of the basic results. In none of these three cases have we provided a definitive analysis and there is plenty of room for contributions to the debate from other authors.

8. Country and Regional Abbreviations

Country and region names have been abbreviated in some figures using the International Standards Organisation (ISO) three-digit alphanumeric codes as the following:

Albania	ALB	Kuwait	KWT
Argentina	ARG	Latvia	LVA
Australia	AUS	Liechtenstein	LIE
Austria	AUT	Lithuania	LTU
Belgium	BEL	Luxembourg	LUX
Belize	BLZ	Macedonia, Rep. of	MKD
Brazil	BRA	Malaysia	MYS
Bulgaria	BGR	Mexico	MEX
Canada	CAN	Moldova, Rep. of	MDA
Chile	CHL	Morocco	MAR
Chinese Taipei	TWN	Netherlands	NLD
Colombia	COL	New Zealand	NZL
Cyprus	CYP	Norway	NOR
Czech Republic	CZE	Peru	PER
Denmark	DNK	Philippines	PHL
England *	ENG	Poland	POL
Finland	FIN	Portugal	PRT
France	FRA	Romania	ROU
Germany	DEU	Russian Federation	RUS
Greece	GRC	Singapore	SGP
Hong Kong	HKG	Slovak	SVK
Hungary	HUN	Slovenia	SVN
Iceland	ISL	South Africa	ZAF
Indonesia	IDN	Spain	ESP
Iran, Islamic Rep. of	IRN	Sweden	SWE
Ireland	IRL	Switzerland	CHE
Israel	IRS	Thailand	THA
Italy	ITA	Tunisia	TUN
Japan	JPN	Turkey	TUR
Jordan	JOR	United Kingdom	GBR
Korea	KOR	United States	USA

* The abbreviation was created by the author.

9. List of Figures and Tables

Figure 1. Old v. rescaled Q50, TIMSS 1995 8th-grade maths
Figure 2. Old v. rescaled Q5, TIMSS 1995 8th-grade maths
Figure 3. Old v. rescaled Q95, TIMSS 1995 8th-grade maths
Figure 4. Old v. rescaled Q95-Q5, TIMSS 1995 8th-grade maths
Figure 5. Q95-Q5 v. Q50, old TIMSS 1995 8th-grade maths
Figure 6. Q95-Q5 v. Q50, rescaled TIMSS 1995 8th-grade maths
Figure 7. Benchmarks and levels
Figure 8. Distribution of scores, PISA reading
Figure 9. Q95-Q5 v. Q50, TIMSS maths
Figure 10. Q95-Q5 v. Q50, PISA reading
Figure 11. Q50-Q5 v. Q95-Q50, TIMSS maths
Figure 12. Kernel density plots for 7th and 8th-grades
Figure 13. Kernel density plots for 3rd and 4th-grades
Figure 14. Average ranks, TIMSS and PISA

Table 1. Estimates of standard errors of the mean, PISA reading
Table 2. Estimates of standard errors of regression coefficients, TIMSS 1999 maths
Table 3. Significance tests, TIMSS maths standard deviation
Table 4. Correlation matrices, TIMSS and IALS (aged 16-25)
Table 5. Traffic lights, TIMSS and IALS (aged 16-25)

10. References

R. Adams and M. Wu (eds.) (2002) *PISA 2000 Technical Report*, OECD, Paris.

A. B. Atkinson (1998), *Poverty in Europe*, Basil Blackwell, Oxford.

C. Beach and R. Davidson (1983), 'Distribution-Free Statistical Inference with Lorenze Curves and Income Shares', *The Review of Economic Studies*, 50(4):723-735.

A. Beaton (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.

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

V. Gallina (ed.) (2000), *La competenza alfabetica in Italia*, CEDE, Rome, and Franco Angeli, Milan.

E. Gonzalez, J. Miles (editors) (2001) *User Guide for the TIMSS 1999 International Database*, Boston College.

International Adult Literacy Survey Microdata User's Guide, Statistics Canada.

M. Kendall and A. Stuart (1969), *The Advanced Theory of Statistics*, Griffin, London.

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

M. Martin, K. Gregory, S. Stemler (eds.) (2000b), *TIMSS 1999 Technical Report*, Boston College.

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

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

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

OECD (2002), *Manual for the PISA 2000 Database*, OECD, Paris.

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

OECD and UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow – Further Results from Pisa 2000*, Paris 2000

UNICEF (2002) *A League Table of Educational Disadvantage in Rich Nations*, Innocenti Report Card 4.

K. Yamamoto and E. Kulick (2000), 'Scaling Methodology and Procedures for TIMSS Mathematics and Sciences Scales' in Martin et al (eds.) (2000b).

Appendix A

Recent and forthcoming cross-national surveys of learning achievement and functional literacy

Survey	Age group	Subjects covered	Results published
Trends in International Maths and Science Study (TIMSS), 1995	10, 14, 18	maths and science	1998
Trends in International Maths and Science Study (TIMSS), 1999	14	maths and science	2000
International Adult Literacy Survey (IALS), 1994-98	16-59	document, prose and quantitative literacy	1996-2000
Programme of International Student Assessment (PISA), 2000	15	reading, maths and science (emphasis on reading)	2001
Programme of International Student Assessment 'Plus' (PISA+), 2002 (PISA extended to non-OECD countries)	15	reading, maths and science (emphasis on reading)	2003
Progress in International Reading Literacy Study (PIRLS), 2001	10	Reading	2003
Adult Literacy and Life Skills Survey (ALLS), 2003 (wave 1)	16-59	document, prose and quantitative literacy	2004
Programme of International Student Assessment (PISA), 2003	15	reading, maths and science (emphasis on maths)	2004 (December)
Trends in International Maths and Science Study (TIMSS), 2003	10, 14	maths and science	2004 (December)

Notes

TIMSS and PIRLS are organised by the International Study Center, Boston College, USA. PISA is organised by OECD. PISA+ is organised by OECD and UNESCO Institute for Statistics. IALS was organised by OECD and Statistics Canada. ALLS is organised by a consortium led by Statistics Canada.

Appendix B

Which countries participated in which surveys

	IALS	3rd/4th grade TIMSS 1995	7th/8th grade TIMSS 1995	8th grade TIMSS 1999	Age 15 PISA 2000/2	4th grade PIRLS
Albania					X	
Argentina					X	X
Australia	X	X	X	X	X	
Austria		X	X		X	
Belgium	X[1]		X	X[1]	X	
Belize						X
Brazil					X	
Bulgaria			X	X	X	X
Canada	X	X	X	X	X	X[2]
Chile	X			X	X	
Chinese Taipei				X		
Colombia			X			X
Cyprus		X	X	X		X
Czech Republic	X	X	X	X	X	X
Denmark	X		X		X	
Finland	X			X	X	
France	X		X		X	X
Germany	X		X		X	X
Greece		X	X		X	X
Hong Kong		X	X	X	X	X
Hungary	X	X	X	X	X	X
Iceland		X	X		X	X
Indonesia				X	X	
Iran, Islamic Rep. of		X	X	X		X
Ireland	X	X	X		X	
Israel		X	X	X	X	X
Italy	X	X	X	X	X	X
Japan		X	X	X	X	
Jordan				X		
Korea		X	X	X	X	
Kuwait		X	X			X
Latvia		X	X	X	X	X
Liechtenstein					X	
Lithuania			X	X		X
Luxembourg					X	
Macedonia, Rep. of				X	X	X
Malaysia				X		
Mexico					X	
Moldova, Rep. of				X		X
Morocco				X		X
Netherlands	X	X	X	X	X	X
New Zealand	X	X	X	X	X	X
Norway	X[3]	X	X		X	X
Peru					X	
Philippines				X		

	IALS	3rd/4th grade TIMSS 1995	7th/8th grade TIMSS 1995	8th grade TIMSS 1999	Age 15 PISA 2000/2	4th grade PIRLS
Poland	X				X	
Portugal	X	X	X		X	
Romania			X	X	X	X
Russian Federation			X	X	X	X
Singapore		X	X	X		X
Slovakia			X	X		X
Slovenia	X	X	X	X		X
South Africa			X	X		
Spain			X		X	
Sweden	X		X		X	X
Switzerland	X		X		X	
Thailand		X	X	X	X	
Tunisia				X		
Turkey				X		X
United Kingdom	X[4]	X[5]	X[5]	X[6]	X	X[5]
United States	X	X	X	X	X	X

1 Belgium is represented by the province of Flanders only in IALS and Flanders plus Brussels in TIMSS 1999 (the two making up the Flemish community of Belgium).

2 Canada is represented by the provinces of Ontario and Quebec only.

3 Norway is represented by Norway Bokmal only.

4 UK is represented by Great Britain and Northern Ireland separately.

5 UK is represented by Scotland and England only.

6 UK is represented by England only.