Investigating the relationship between aspects of countries' assessment systems and achievement on the Programme for International Student Assessment (PISA) tests

Tim Gill and Tom Benton

Cambridge Assessment Research Report

30th March 2013

University of Cambridge Local Examinations Syndicate

Author contact details:
Tim Gill
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
Gill.T@cambridgeassessment.org.uk

Tom Benton
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
Benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

How to cite this publication:

Gill, T., and Benton, T. (2013). *Investigating the relationship between aspects of countries' assessment systems and achievement on the Programme for International Student Assessment (PISA) tests*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Contents

Executive summary

The main aim of this research was to explore the association between attainment of pupils in PISA 2009 and details of assessment systems within different countries; including the existence of national testing, the aims of assessments and the ways in which results are used. In particular, the aim was to examine whether there are any aspects of assessment systems that are likely to be beneficial to students in the UK.

Analysis was undertaken in three stages. To begin with the extent of the differences in the background characteristics of candidates in different countries was assessed. This enabled us to identify a number of countries deemed sufficiently similar to the UK for statistical comparisons to be meaningful. Having restricted ourselves to countries deemed sufficiently similar, the relative achievement of these countries was compared, but, crucially, after making statistical adjustments to account for the influence of student background. Finally, multilevel modelling was used to explore the relationship between the details of assessment systems and the achievement of students once the impact of background characteristics is accounted for.

The main findings of the research are:

- Using data from international studies to explore the potential for elements of an education system to provide improvement in a UK context is not straightforward. We found that for many countries, including (disappointingly) many of the high profile top performers, the characteristics of students were too different from those of their UK counterparts for any meaningful comparisons to be made. By restricting our attention to those countries most similar to the UK we improve our chances of making conclusions from the data that are relevant within our own context.
- The international rankings of countries are to some extent driven by the background characteristics of their students. Once these characteristics are accounted for, there can be substantial changes in the estimated relative performance of countries. Indeed, the main method used to make adjustments in this report indicated a substantially improved ranking for the UK, although alternative methods of adjustment may not necessarily yield the same result.
- The evidence for the details of assessment systems being a major driver of improvement in educational systems is hardly overwhelming. Very few statistically significant effects were identified relating the details of countries' assessment systems to their performance in PISA 2009.
- Some evidence was found suggesting that using test results to monitor and evaluate teachers may have a negative impact on student attainment. However, no negative association was found between the extent to which schools are required to publish their results (for example in league tables) and the overall performance of a country.

Overall this report highlights the difficulties involved in relating performance in international tests to specific aspects of educational systems. While it is tempting to examine the characteristics of education systems in high performing jurisdictions and hope that translating the systems from these countries will also lead to improved performance, such an approach ignores two crucial issues; whether candidates in one country have anything in common with candidates in our own country, and whether the highlighted aspects of the educational system are unique to high performing countries or whether they can also be found in those with low performance. It is only when we take account of all of the data and

adjust for other influential factors that we can get a true picture of the influence of a system level variable; albeit an inconclusive one.

Introduction

International benchmarking studies such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) provide a rich source of data regarding the relative effectiveness of education in different countries. The data is publicly available but is rarely analysed in any detail by UK researchers. The aim of this research was to use this data to explore the extent of the link between the assessment systems within countries and their overall level of proficiency in reading, maths and science. In analysing this data it was also possible to compare the performance of the countries participating in PISA, after accounting for a number of background variables of the students taking part.

Background and literature review

The availability of data from international studies of student achievement has generated a wealth of research comparing countries' performance and attempting to determine what factors might be important in raising achievement (see Hanushek and Wößmann, 2011 for a review). One strand of this focuses on aspects of the assessment systems in the different countries and, in particular, the existence or not of external exit exams. There are two ways of looking at the effect of these types of exams; between countries at a particular point in time, and over time, following their introduction into countries' assessment systems. For the purpose of this report we focus on research using the first of these methodologies.

Almost all of this literature suggests that countries with external exit exams tend to have higher levels of achievement, as measured by international tests, than those that do not. For example, Bishop (1997) compared the performance of students in countries taking part in the TIMSS assessment in 1995, by whether or not they had Curriculum Based External Exit Exams (CBEEEs) in secondary schools, whilst taking account of GDP per capita. He found a significant effect on median test score in both maths and science of being in a country with CBEEEs, equivalent to one grade[1] level in maths and 1.2 grade levels in science. Wößmann (2003) used the same data from the 1995 TIMSS to assess the impact of a number of 'institutional features' of a country (including central exams) on performance. Controlling for the effects of family background and resources spent on education students in countries with central exams scored on average 16.1 points[2] higher in maths and 10.7 points higher in science (both differences were statistically significant).

Results of a similar magnitude using data from PISA in 2000 were found by Fuchs and Wößmann (2007). Students in countries with external exit exams performed better by 19.1 points in maths (statistically significant) and 15 points in science (not statistically significant). As before, this is after controlling for student background and resource endowment. OECD (2007) used data from PISA 2006 in science and found that students in countries with external exit exams performed better by an average of 17 points (after accounting for socio-economic and demographic factors). However, this difference was not statistically significant. Similarly, OECD (2010) found an average difference of 16 points in reading score in PISA 2009 between students in countries with external exit exams and those without. This was statistically significant, although only at the 7% level.

[1] 'Grade' in this context refers to school year.
[2] TIMSS is scaled to have a mean of 500 and a standard deviation of 100 points

Mons (2009) review of international research into external exit exams and student performance concludes that the evidence of a link is inconsistent. She refers to her own work from 2007 (no ref) which looked at the same PISA 2000 data as Fuchs and Wößmann (2007) but, crucially, accounted for the level of economic development in the country. When controlling for GDP per capita she found no significant effect of centralised exams on student attainment.

Jurges and Schneider (2004) found a similar lack of a significant effect when analysing TIMSS data from 1995. They essentially reproduced Bishop's analysis of the data, but only for OECD countries (excluding less economically developed 'partner' countries). Thus the effect of external exit exams seems to be less when considering more economically developed OECD countries only.

Wößmann (2005) uses data from TIMSS in 1995 and 1997 and PISA in 2000 to assess whether the effect of central exit exams was different across other factors. He found evidence in TIMSS 1995 science and maths and PISA that the effect of central exit exams was greater for higher ability students. In TIMSS there was also a greater effect for immigrants and for those living with both parents. Finally, he found that the effect of external exit exams increases over grade levels, being 17.5% of a standard deviation higher in the eighth grade than in the seventh grade (for TIMSS maths performance).

Many of these studies acknowledge that, despite taking account of student and school characteristics, it is necessary to be cautious when comparing countries that do and do not have central exit exams because their performance in international tests can be influenced by many other, unobserved, factors. For example, countries with CBEEEs may place a higher priority on education generally (Jurges et al, 2003) or there could be substantial differences in the cultural and historical context between countries (Fuchs and Wößmann, 2007; Wößmann, 2003).

Two studies attempt to overcome these issues by investigating the effect of central exit exams *within* a country by comparing performance between federal states with differing assessment systems. Bishop (1997) compared students' performance on the 1991 IAEP (International Assessment of Educational Progress) in Canadian provinces with and without curriculum-based external examinations. Taking into account background variables such as books in the home, language spoken at home, levels of parental involvement and hours spent doing homework, the existence of external exit exams had a large positive impact; almost ¼ of a standard deviation (4/5 of a grade level) in maths and 17% of a standard deviation (3/5 of a grade level) in science. Jurges et al (2003) took advantage of different assessment systems in federal states in Germany. Using data from TIMSS 1995 they found a difference in maths performance of about 1.25 grade equivalents between students in states with external exit exams and those without (after taking into account background variables). To further account for unobserved differences between states they also looked at differential performance in maths and science of students in states with external exit exams in maths, but not in science. Students in states with external exit exams in maths only had a better relative performance in maths than in science, by about 1/3 of a grade equivalent.

Although the existence of central exit exams is seen by many to be the aspect of assessment systems most likely to impact on student performance, other aspects have also been investigated. In particular, OECD (2007, 2010) used data taken from PISA 2006 and 2009 to look at the effect of performance data being reported to parents, used to monitor and

evaluate teachers or allocate resources, tracked over time by an administrative authority and posted publicly. The only one of these factors that had a statistically significant effect was posting achievement data publicly, which was positively related to student performance in the 2006 data. OECD (2010) also reported an interaction effect between posting achievement data publicly and levels of school autonomy in allocating resources. In countries where no schools posted achievement data, student performance was significantly negatively related to levels of autonomy, whereas in countries where all countries posted achievement data there was a significant positive relationship. However, in both cases the effects were really quite small, and only just reached statistical significance.

OECD (2010) also investigated the effect of regular use of standardised tests in schools on student attainment, but found no statistically significant relationship. However, an interesting interaction was found by Fuchs and Wößmann (2007) in the PISA 2000 data, with the use of standardised testing at least once a year positively related to performance in countries with external exams and negatively related to performance in countries without external exams. This effect was consistent across all three subjects, being worth about 5-7 points in countries with external exams.

Thus there is some evidence that countries with external exit exams tend to perform better than those without. However, doubts still remain that this link is causal, given the difficulty of comparing performance between countries which differ in many other areas and with some studies suggesting that this relationship disappears when GDP is taken into account. Other aspects of assessment systems tend to have little or no effect.

This research takes a new approach to the issue by attempting to limit comparisons to countries that are sufficiently similar to the UK in terms of their background characteristics for such comparisons to be meaningful. In this way we will explore the extent to which the internationally published evidence on the potentially positive effects of an assessment system might be applicable within a UK context. Thus, the first part of the research is to create an indicator for each country of similarity of their students to UK students. Once similar countries have been identified, their performance in PISA 2009 is compared, after accounting for background variables. Finally, multilevel modelling is undertaken to investigate the relationship between countries' assessment systems and their performance on PISA.

The research questions are outlined below:

1) *To what extent can the performance of UK students be compared to that of other countries? Are students sufficiently similar in terms of their background characteristics for such comparisons to be meaningful?*

2) *How does the performance of the UK compare to that of other countries once all relevant background characteristics are taken into account?*

3) *To what extent is the relative performance of comparable countries associated with the assessment systems that are used?*

Data

The Programme for International Student Assessment (PISA) is a study of the educational levels in OECD and 'partner' countries around the world, as measured by the skills and knowledge of their 15 year olds. Every three years since 2000 a group of randomly selected students from each country has taken tests in reading, maths and science. Students and school principals also complete background questionnaires providing information on family background and school policies and practices. The results of the tests are summarised at the national level and countries are then ranked without any attempt to take account of the students' backgrounds, school policies or the educational system in the country as a whole. However, the data is also made available to researchers at individual student and school level.

At the time of this research, the latest year for which data was available was 2009. The performance data and responses to the questionnaires for that year were downloaded from the PISA website (http://pisa2009.acer.edu.au/). The full performance database consists of the results for 519,958 students in 18,641 schools in 74 countries. However, for the purpose of this research this database was reduced to include only countries sufficiently similar to the UK for meaningful comparisons to be made (see research question 1, below). The performance database also includes the responses to the student questionnaire. A separate database of responses to the school questionnaire was also downloaded.

The main area of interest for this research was the relationship between a country's assessment system and their performance on international tests. Some of the questions in the school questionnaire relate to assessment policy and practice and responses to these were used in the analysis. However, it is also of interest to investigate the impact of assessment systems at a national level; hence, alternative sources of data were investigated. Whilst looking through PISA reports a table was found listing all OECD and partner countries and whether or not they had standards-based external exams (OECD, 2010, p229, table IV.3.11). A second source of data were reports on assessment systems of countries in the EU (EACEA, 2009a; EACEA, 2009b), which provided information on a number of aspects related to assessment systems.

Research question 1: To what extent can the performance of UK students be compared to that of other countries? Are students sufficiently similar in terms of their background characteristics for such comparisons to be meaningful?

Method

The first part of this research attempted to identify countries similar to the UK in terms of their background characteristics. As mentioned in the introduction, one issue with trying to make comparisons between countries on their performance in international tests is trying to account for differences in their economic, social and cultural contexts (Crisp, 2013). One way of overcoming this issue is to reduce these differences by only including countries that are reasonably similar.

The method for determining which countries were most similar to the UK involved using data from the PISA student questionnaire to calculate an indicator for each student in non-UK

countries of the extent of their similarity to UK students in terms of background variables. To do this a logistic regression was run, with being a UK student as the dependent variable and a selection of background factors as independent variables. There are a large number of potential variables in the PISA student questionnaire so to reduce processing time it was decided to select the 10 most significant predictors of PISA scores. These were selected by including all variables related to student background as independent variables in a regression model with mean PISA score across all subjects and all plausible values as the dependent variable[3]. The least significant predictor (as measured by the sum of squares) was then removed, and the model re-run. This process continued, removing the least significant predictor each time until only 10 variables remained. These were:

- *Family structure (single parent, two parents, other)*
- *What language do you speak at home most of the time (language of test, another language)?*
- *Which of the following are in your home?*
    - *A link to the internet;*
    - *Classic literature;*
    - *Books of poetry;*
    - *A dictionary.*
- *How many of these are there at your home?*
    - *Computers;*
    - *Cars.*
- *How many books are there in your home?*
- *Highest parental occupational status (continuous variable from 16 (low socio-economic status) to 90 (high socio-economic status))*

The output from the logistic regression model included a predicted probability for each student of 'being a UK student', given his or her values of the independent variables. This probability was denoted as their 'propensity' score (i.e. their propensity for being from the UK). A weight was then calculated of the likelihood of being a UK student (weight = propensity score / (1-propensity score)). Thus, a student with a predicted probability of being from the UK of 0.7 would have a weight of 0.7 / 0.3 = 2.33. This indicates that they were two and a third times more likely to be a UK student than not.

Provided the logistic regression model fits the data, then the weights calculated in this way can be applied to account for the differences between countries. An example of this is shown in Figure 1. This shows the difference between the UK and Australia on each of the variables listed above before and after weights are applied as described. For the purposes of brevity, only one category from each of the multi-category variables (such as the number of books in the home) are shown in this chart (although all the categories were used in weighting). Also, because highest parental occupational status is a continuous variable, it is not suitable to be displayed in this format and so is not included. The chart shows that before weighting there was a reasonable degree of similarity between the UK and Australia across many of the background variables. However, some notable differences were evident in terms of the

[3] This approach is not methodologically pure. Both averaging over plausible values and subjects, and using ordinary linear regression rather than a method accounting for the hierarchical structure of the data are likely to lead to an incorrect estimation of the standard errors of the model. However, since the purpose of this preliminary analysis is only to identify the 10 variables for which it is most crucial that there is overlap in terms of the background characteristics of students across different nations, the methodology is sufficient.

numbers of computers, cars, and books in each student's home. Once the weights have been applied, the differences between the UK and Australia are largely corrected. This would imply that weighted achievement data from Australia can be compared to the achievement of UK students on a like-with-like basis.

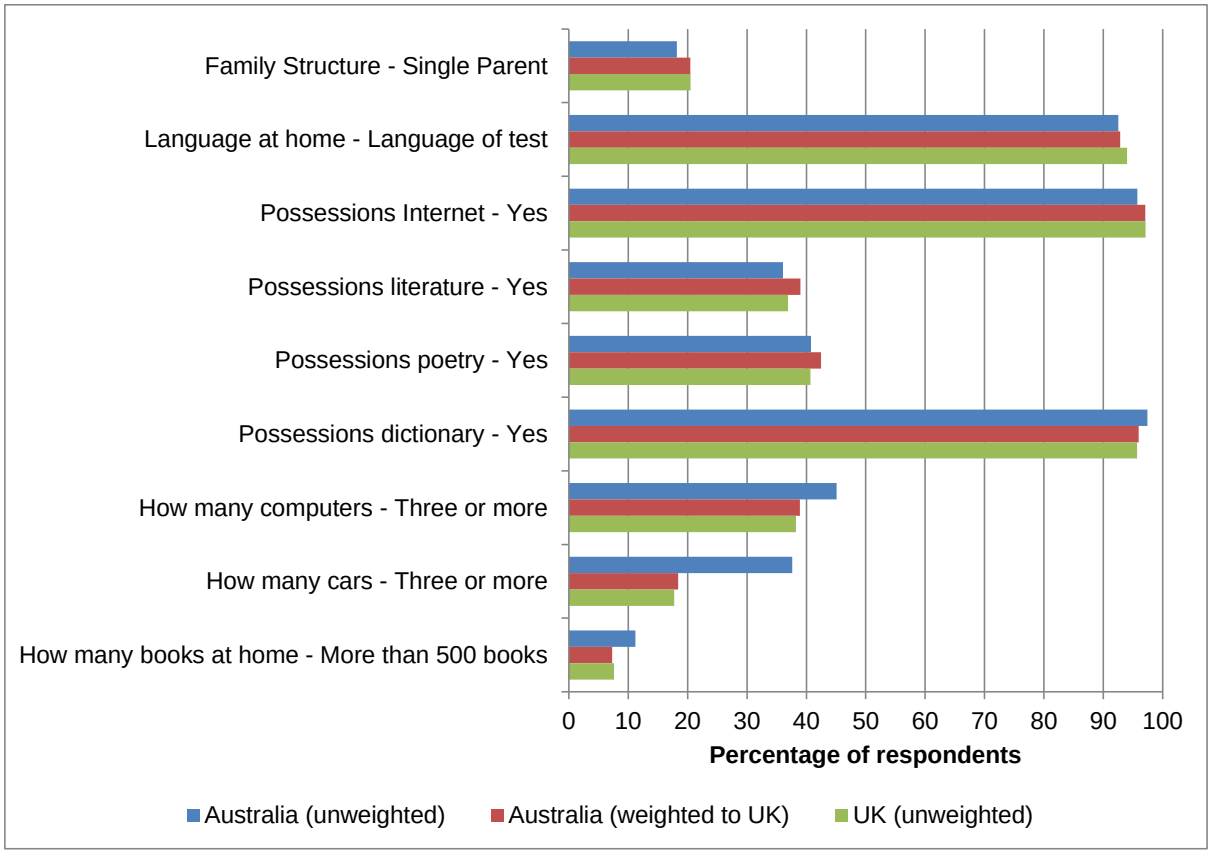


Figure 1: Example of using propensity score based weights to adjust for differences between UK and Australian students

In using a method such as this to account for differences between countries we are implicitly assuming that the background variables we have controlled for have a similar meaning across countries. That is, for example, that having three cars in your home has the same sort of implications in Australia in terms of the socio-economic background of different students as it does in the UK. This is a less than ideal assumption. However, we cannot avoid such assumptions without abandoning all hopes of comparing countries on a like-with-like basis. Indeed, the same kinds of assumptions (if not involving exactly the same variables) are also present in much of the OECD's own analyses of PISA data. For example, the OECD's analysis of the interaction of the effects of school accountability and autonomy on performance (OECD 2010, page 171, Table IV.2.5) implicitly assumes a common meaning across countries for school autonomy, private schooling, the index of economic, social and cultural status[4], school size and school location. For our own analysis, all we can do is note that these assumptions (that background variables have a common meaning across countries) are an unavoidable caveat on the validity of our analyses, and agree with other authors that making meaningful international comparisons remains a far from straightforward undertaking.

[4] Even though the definition of this index itself varies across countries.

Analysis of the weights created via propensity scores provided a means by which it was possible to quantify the similarity between the background characteristics of students in the UK and the characteristics of students in each other country. This calculation was done by means of *effective sample sizes*. The effective sample size for any country tells us how much the data from that country is worth once it has been adjusted to make the background characteristics of its students equivalent to those in the UK. A technical explanation of how effective sample sizes have been calculated is given in Appendix A and the results of this analysis are shown in Table 1. This shows that, for example, data from 5,509 students was available from Finland but the effective sample size for this country is 3,315. This means that, although data from over five thousand students has been collected from Finland, once we have applied our weights, any estimates based on this data will only be as accurate as those we would gain from a sample of just under three and a half thousand students had these students been specifically picked to match the characteristics of students in the UK.

Calculating effective sample sizes as a percentage of the actual number of respondents in any country provides a means by which the similarity of each country to the UK can be quantified. For example, making the characteristics of Finnish students match the characteristics of those in the UK is equivalent to discarding data from around 40 per cent of Finnish respondents. However, making the characteristics of Brazilian students match with their UK counterparts requires a degree of weighting equivalent to discarding data from almost 95 per cent of respondents. Thus we conclude that students from the UK are somewhat more similar to those in Finland than to those in Brazil.

Once the effective sample sizes were calculated they were used to rank the countries, with the largest effective sample size (as a percentage of actual sample size) indicating the country with students most similar to UK students. It was then necessary to decide a cut-off point, below which were countries with students deemed too dissimilar to be included. This decision was somewhat arbitrary, based on wanting to ensure a reasonably large number of countries to be included in the comparisons but without including countries very dissimilar to the UK. The final decision was to include all countries with an effective sample size greater than 20% of the actual sample size.

Results

The results of this analysis are shown in Table 1. There were 34 countries which were above the cut-off point. Inspection of the table suggests this was a reasonable result as it meant that the countries retained were mostly in Western Europe or were other English-speaking countries, whilst the excluded countries were mostly in Eastern Europe, the Middle East, the Far East and South America and would therefore be expected to be rather dissimilar to the UK.

For the remaining analyses only the 35 countries (34 + UK) in this reduced dataset were included, unless otherwise stated.

Table 1: Effective sample sizes for all PISA countries

Country	Actual Sample size	Effective sample size (n)	Effective sample size (% of actual)
Liechtenstein	276	212.9	77.1%
Finland	5,509	3,314.5	60.2%
Belgium	7,105	4,118.4	58.0%
Ireland	3,432	1,876.8	54.7%
Germany	3,802	2,004.1	52.7%
Sweden	3,975	1,926.6	48.5%
Austria	5,443	2,628.1	48.3%
Canada	20,459	9,263.8	45.3%
Australia	12,413	5,285.7	42.6%
Netherlands	4,170	1,755.8	42.1%
Israel	4,333	1,784.0	41.2%
New Zealand	4,221	1,721.9	40.8%
France	3,743	1,516.2	40.5%
Switzerland	10,422	4,174.5	40.1%
Iceland	3,370	1,285.7	38.2%
Norway	4,291	1,616.5	37.7%
Portugal	5,844	2,070.4	35.4%
United States	4,735	1,673.9	35.4%
Miranda-Venezuela	2,196	723.0	32.9%
Slovenia	5,241	1,705.9	32.5%
Czech Republic	5,428	1,700.0	31.3%
Denmark	4,834	1,456.5	30.1%
Estonia	4,422	1,255.5	28.4%
Latvia	4,007	1,016.3	25.4%
Croatia	4,485	1,130.2	25.2%
Chinese Taipei	4,996	1,238.9	24.8%
Slovak Republic	4,135	1,000.3	24.2%
Trinidad and Tobago	3,713	864.6	23.3%
Spain	23,706	5,326.5	22.5%
Bulgaria	3,609	810.2	22.5%
Hungary	4,228	949.0	22.4%
Greece	4,479	1,000.3	22.3%
Poland	4,535	1,004.6	22.2%
Lithuania	4,011	878.5	21.9%
Italy	26,313	4,929.8	18.7%
Macao-China	5,412	1,013.7	18.7%
Argentina	3,841	696.8	18.1%
Chile	5,079	889.8	17.5%
Montenegro	3,905	677.0	17.3%
Panama	2,599	447.8	17.2%
Japan	5,323	903.8	17.0%
Korea	4,768	790.5	16.6%
Costa Rica	3,655	571.4	15.6%
Uruguay	4,879	755.3	15.5%
Thailand	4,943	761.8	15.4%
Malaysia	4,034	612.2	15.2%
Jordan	4,783	695.4	14.5%
Serbia	4,893	689.2	14.1%
Qatar	5,696	772.9	13.6%
Malta	2,933	396.4	13.5%
Singapore	4,855	644.1	13.3%
United Arab Emirates	7,885	1,017.0	12.9%
Romania	4,149	483.0	11.6%
Russian Federation	4,742	549.3	11.6%
Republic of Moldova	3,909	436.0	11.2%
Georgia	2,884	318.1	11.0%
Luxembourg	3,797	413.8	10.9%
Hong Kong-China	4,537	458.1	10.1%
Tunisia	3,809	357.5	9.4%
Turkey	3,832	356.1	9.3%
Shanghai-China	4,982	455.4	9.1%
Albania	3,509	298.6	8.5%
Azerbaijan	3,113	248.7	8.0%
Colombia	6,297	460.1	7.3%
Kazakhstan	4,880	354.1	7.3%
Brazil	15,818	954.5	6.0%
Peru	5,199	312.7	6.0%
Himachal Pradesh-India	707	40.8	5.8%
Mexico	32,253	1723.3	5.3%
Kyrgyzstan	3,526	169.4	4.8%
Indonesia	3,691	158.9	4.3%
Tamil Nadu-India	1,774	69.2	3.9%
Mauritius	3,721	129.3	3.5%

Table 1 highlights the difficulty involved with making international comparisons on a like-with-like basis. It can immediately be seen that, even when we only try and match across the 10 background characteristics listed earlier, internationally there are no countries which provide unproblematic comparisons with the UK. Indeed there are only five countries (Liechtenstein, Finland, Belgium, Ireland and Germany) where the weighting required to make the student populations equivalent is comparable to discarding less than half of the respondents from these countries. In contrast, there are 39 countries where the required weighting is equivalent to discarding more than 80 per cent of the data. This includes many high profile Asian countries such as Hong Kong, Korea and Singapore and implies that any kind of comparison with the UK that takes account of the differences in the background characteristics of students is just not possible with any meaningful degree of accuracy.

To check whether the effective sample size method was producing reasonable results the responses to the questions used in the analysis were compared for the UK and several of the other countries. These were Finland (similar and commonly discussed), Netherlands (fairly similar and commonly discussed), Italy (as it just misses out on being included in comparisons), Singapore (commonly discussed but nothing like the UK) and Shanghai China (also commonly discussed but nothing like the UK). The results are shown in Figure 2.

The results seem to be consistent with the outcome of the effective sample size analysis. Finland is very similar to the UK on most of the variables, as is the Netherlands. Shanghai is clearly very dissimilar to the UK on all indicators except for the final two. Italy is dissimilar on most of the indicators, but not to a large degree on any of them. The results for Singapore are less clear, with the UK seemingly quite similar on a number of the indicators. However, where the differences do occur they are large (e.g. language spoken at home, number of cars at home), or relatively large (e.g. dictionary at home, with five times fewer students in Singapore without a dictionary compared with the UK (1% and 4.9% respectively)). This suggests the outcome of the effective sample size analysis was reasonable.

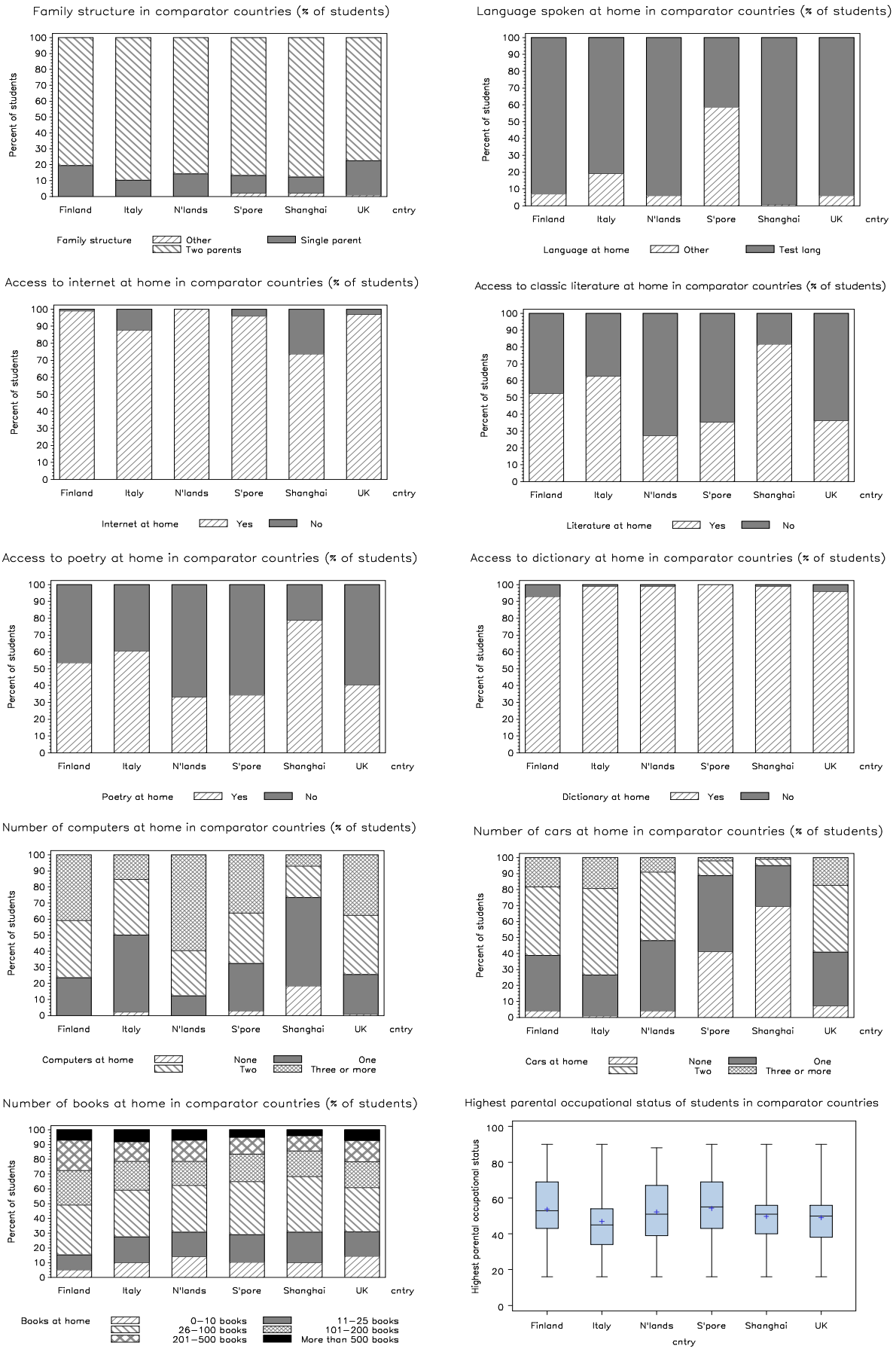


Figure 2: Comparison of responses to questions used in effective sample size analysis

Research question 2: How does the performance of the UK compare to that of other countries once all relevant background characteristics are taken into account?

Method

To answer this question a series of multilevel models were run to look at the performance on PISA by students in each country, after accounting for a number of background variables. Multilevel models are useful in this context because they recognise that the 'clustering' of individuals within schools means that students within a school are likely to have more in common with each other than with students in different schools. Thus, the models had two levels (students nested in schools).

The dependent variable was the performance of students on each of the PISA tests. An important aspect of the reporting of scores for individual students on PISA tests is the concept of plausible values. Instead of reporting one score for each student on a test, a probability distribution of their ability is estimated (i.e. a range of possible scores, each with an associated probability). Plausible values are random draws from this probability. The PISA database contains five plausible values for each student on each test. When undertaking data analysis using the plausible values (e.g. modelling) it is necessary to run the analysis separately for each plausible value and then combine the results. For more information on the rationale for plausible values and how they are used see the PISA data analysis manual (OECD, 2009).

Thus, five models were run for each test, one for each plausible value and the parameter estimates from each model were combined to give an overall average parameter estimate. Standard errors of the parameter estimates were calculated by combining information from the standard errors calculated within each individual model and also taking account of the variability in parameter estimates between the five models.

The predictor variables were several student level background variables and the student's country of residence. Adding countries to the model as a fixed effect (rather than as another hierarchical level) allowed estimates to be made of their effectiveness. The country with the largest (positive) parameter estimate is the one that (according to the model) had the largest positive effect on student performance, after accounting for background variables.

The background variables that were used in this analysis were taken from the student questionnaire administered to all students taking the tests. Most were the answers given by students to questionnaire items, but some were derived from answers to more than one item (for example, highest educational level of parents was derived from answers about the educational level of each parent). The variables were:

- *Family structure (single parent, two parents, other)*
- *Highest educational level of parents (ISCED 1 (lowest) – ISCED 5A,6 (highest)).*
- *Immigration status (native, first generation, second generation)*
- *What language do you speak at home most of the time (language of test, another language)?*
- *Which of the following are in your home? A desk to study at; A room of your own; A quiet place to study; A computer you can use for school work; Educational software; A link to the internet; Classic literature; Books of poetry; Works of art; Books to help*

with your school work; Technical reference books; A dictionary; A dishwasher; A DVD player.

- *How many of these are there at your home? Cellular phones; Televisions; Computers; Cars; Rooms with a bath or shower.*
- *How many books are there in your home (0-10, 11-25, 26-100, 101-200, 201-500, 501+)?*
- *Highest parental occupational status*
- *Age (in fractions of years, 2 d.p.)*
- *Index of economic, social and cultural status (PISA derived index, mean of 0 and sd of 1)*[5]

Results

Only countries that were deemed sufficiently similar to the UK were included in these models (see Table 1). Tables 2 to 4 present the parameter estimates for each country for each of the three PISA measures (maths, reading and science). These tables also include the original PISA scores and the rankings for each country, recalculated for the countries included in this analysis only. For example, the highest ranked country in the PISA reading test was Shanghai-China but this country was not included. Finland was the highest ranked country that was included, so is given a ranking of 1 in the table.

It should be noted that Miranda-Venezuela undertook their tests in 2010, rather than 2009 and as such they did not appear in the published rankings of PISA scores. However, it was possible to calculate their scores manually and these were used in the recalculated rankings.

The country parameter estimates are the change in PISA scores attributable to being in that country in comparison to the UK, whilst holding constant the background variables. The shading in the tables represents the level of statistical significance of the parameter estimates. The lightest shading indicates that the parameter estimate was not statistically significant, meaning that after accounting for background variables there was no real difference in PISA scores in that country compared to the UK. The next heaviest level of shading indicates statistical significance at the 5% level, whilst the heaviest shading indicates statistical significance at the 1% level.

Thus the most effective country in terms of achievement on PISA was New Zealand in reading and science and Chinese Taipei in maths. Comparing the rankings with the rankings according to the PISA score shows that there was some similarity between the two, and there seemed to be more similarity amongst countries towards the bottom of both rankings. However, some countries had rankings according to the model presented here that were notably higher than their PISA rankings, including Portugal (improved by 11, 8 and 10 places in reading, maths and science respectively), United Kingdom (12, 9 and 4) and Ireland (9, 5 and 6). Countries with notably lower rankings according to the models include Slovenia (-7, -15 and -14), Norway (-12, -10 and -11) and Hungary (-8, -5 and -12).

[5] For further information on the calculation of this index see http://www.oecd.org/pisa/pisaproducts/pisa2009/48579081.pdf (p21-22)

Table 2: Country effects and PISA ranking (Reading)

Ranking	Country	Parameter estimate	Ranking according to PISA measure	PISA score
1	New Zealand	19.9	3	521
2	Netherlands	18.2	5	508
3	Canada	15.2	2	524
4	Finland	12.7	1	536
5	Belgium	11.4	6	506
6	Australia	6.9	4	515
7	United States	5.6	12	500
8	Ireland	1.6	17	496
9	United Kingdom	0.0	21	494
10	Germany	-1.5	14	497
11	Portugal	-6.0	22	489
12	Poland	-7.4	11	500
13	Iceland	-9.2	10	500
14	France	-9.8	16	496
15	Sweden	-10.2	15	497
16	Denmark	-12.8	19	495
17	Switzerland	-13.1	9	501
18	Chinese Taipei	-13.2	18	495
19	Norway	-16.2	7	503
20	Estonia	-19.1	8	501
21	Greece	-21.4	24	483
22	Liechtenstein	-23.0	13	499
23	Croatia	-24.5	29	476
24	Czech Republic	-24.8	27	478
25	Spain	-30.1	26	481
26	Israel	-32.3	30	474
27	Slovak Republic	-32.3	28	477
28	Hungary	-32.5	20	494
29	Latvia	-32.5	23	484
30	Austria	-38.2	31	470
31	Lithuania	-43.3	32	468
32	Slovenia	-51.2	25	483
33	Bulgaria	-67.5	33	429
34	Miranda-Venezuela	-78.5	34	422
35	Trinidad and	-84.2	35	416

For the UK, this analysis suggests a better performance than that given by the PISA rankings. The best performance by UK students was in science (7th place), followed by reading (9th) and maths (13th). In science, only three of the countries involved in this comparison performed better than the UK to a statistically significant degree. Only five countries performed better to a statistically significant degree in reading, whilst nine did so in maths.

Table 3: Country effects and PISA ranking (Maths)

Ranking	Country	Parameter estimate	Ranking according to PISA measure	PISA score
1	Chinese Taipei	46.7	1	543
2	Netherlands	35.3	6	526
3	Belgium	25.7	8	515
4	Switzerland	25.4	4	534
5	Finland	24.3	2	541
6	New Zealand	24.2	7	519
7	Liechtenstein	21.3	3	536
8	Canada	19.6	5	527
9	Germany	17.0	10	513
10	Australia	10.5	9	514
11	Estonia	3.7	11	512
12	Slovak Republic	1.7	17	497
13	United Kingdom	0.0	22	492
14	Iceland	-1.2	12	507
15	Czech Republic	-1.4	21	493
16	Poland	-1.9	19	495
17	France	-2.2	16	497
18	Portugal	-2.3	26	487
19	Denmark	-2.5	13	503
20	Ireland	-2.7	25	487
21	United States	-3.9	24	487
22	Austria	-6.8	18	496
23	Sweden	-8.3	20	494
24	Spain	-14.3	27	483
25	Norway	-18.0	15	498
26	Latvia	-19.7	28	482
27	Lithuania	-21.7	29	477
28	Hungary	-22.4	23	490
29	Slovenia	-24.7	14	501
30	Greece	-27.4	30	466
31	Croatia	-30.4	31	460
32	Israel	-50.4	32	447
33	Bulgaria	-59.0	33	428
34	Trinidad and Tobago	-73.0	34	414
35	Miranda-Venezuela	-97.8	35	397

Table 4: Country effects and PISA ranking (Science)

Ranking	Country	Parameter estimate	Ranking according to PISA measure	PISA score
1	New Zealand	16.2	2	532
2	Netherlands	15.2	6	522
3	Finland	10.4	1	554
4	Germany	6.5	9	520
5	Australia	2.9	5	527
6	Canada	1.9	3	529
7	United Kingdom	0.0	11	514
8	Ireland	-2.6	14	508
9	Belgium	-2.7	15	507
10	Chinese Taipei	-3.9	7	520
11	Estonia	-6.7	4	528
12	United States	-10.0	17	502
13	Switzerland	-15.1	10	517
14	Poland	-15.7	13	508
15	Portugal	-17.3	26	493
16	Czech Republic	-17.3	19	500
17	Liechtenstein	-21.6	8	520
18	France	-24.0	21	498
19	Austria	-26.9	24	494
20	Denmark	-28.2	20	499
21	Croatia	-29.5	30	486
22	Sweden	-30.1	23	495
23	Slovak Republic	-31.6	28	490
24	Iceland	-32.7	22	496
25	Lithuania	-34.1	27	491
26	Slovenia	-36.5	12	512
27	Latvia	-36.8	25	494
28	Hungary	-37.4	16	503
29	Norway	-38.3	18	500
30	Spain	-38.7	29	488
31	Greece	-49.7	31	470
32	Israel	-65.9	32	455
33	Bulgaria	-71.9	33	439
34	Miranda-Venezuela	-97.9	35	422
35	Trinidad and Tobago	-104.4	34	410

Research question 3: To what extent is the relative performance of comparable countries associated with the assessment systems that are used?

Method

The main aim of this research was to investigate the effect of assessment systems on PISA scores in countries similar to the UK, after accounting for background variables. Once the comparable countries were identified these were included in a series of multilevel models. This time the models had three levels (students nested in schools, nested in countries). The extra level in these models takes account of the fact that schools within a country are likely to have more in common with each other than with schools in another country, because they have to work in the same education system and are likely to have similar policies and procedures and similar levels of funding.

The variables used to describe the assessment systems in each country were taken from three different sources. First, data was taken from the PISA school questionnaire, where a number of questions related to assessment systems were asked.

Since there were a large number of these questions (19) it was decided to reduce the set of questions to a more manageable number. To do this, for each question the percentage of pupils in the country attending schools giving each response was calculated. These percentages were then correlated with the parameter estimates for each country in each test calculated in the previous section. This gave an indication of the variables most related to country effects. Table 5 presents the correlation coefficients for each of the questions, for each test. Only variables where the correlation coefficient was significant for at least one test were retained. These are highlighted in bold in the table.

Another variable taken from PISA was the existence or not of standards-based external exams. This data was taken from an annex to an OECD report (OCED, 2010, p229), which lists all OECD and partner countries and indicates the extent to which these exams exist in the country. The definition of standards-based external exams used by OECD is that they are exams that measure performance relative to an external standard, not in comparison to other students in the classroom or school. Furthermore, the results of such exams have real consequences for certification or progression in the education system.

Most countries in the OECD data had a figure of 0 (no standards-based external exams in secondary schools) or 1 (all standards-based external exams). Countries with values between 0 and 1 had standards-based external exams in some part of the system (e.g. in some regions, or some educational programmes). In this table there was no data for France, but another document (EACEA, 2009b, p239) provided the required information. Miranda-Venezuela was also missing from this table, so was excluded from these particular models.

Table 5: Correlation coefficients between country parameter estimates and percent of students in schools giving each response

Question		Reading	Maths	Science
Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using standardised tests?	Never	-0.06	0.05	0.06
	At least 1-2 times a year	0.06	-0.05	-0.06
	At least 3-5 times a year	-0.19	-0.19	-0.30
	At least monthly	-0.20	-0.20	-0.24
	More than once a month	-0.19	-0.30	-0.28
Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using teacher-developed tests?	Never	-0.10	0.08	0.01
	At least 1-2 times a year	0.10	-0.08	-0.01
	At least 3-5 times a year	0.08	0.04	0.02
	At least monthly	0.10	0.15	0.02
	More than once a month	0.14	0.20	0.10
Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using student portfolios?	Never	-0.11	-0.17	-0.10
	At least 1-2 times a year	0.11	0.17	0.10
	At least 3-5 times a year	0.05	0.05	-0.03
	At least monthly	-0.04	<0.01	-0.11
	More than once a month	-0.08	0.03	-0.13
Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using student assignments / projects / homework?	Never	-0.08	-0.10	-0.18
	At least 1-2 times a year	0.08	0.10	0.18
	At least 3-5 times a year	0.12	0.08	0.18
	At least monthly	0.03	0.03	0.09
	More than once a month	0.03	0.04	0.09
In your school, are assessments of students in <national modal grade for 15-year-olds> used for any of the following purposes?	Make decisions about student's retention	-0.06	<0.01	0.10
	Group students for instructional purposes	0.10	-0.09	0.02
	Compare the school to district /national performance	0.18	0.04	0.19
	Monitor school's progress	-0.19	**-0.38**	-0.21
	Make judgments about teacher effectiveness	**-0.42**	**-0.42**	-0.33
	Identify aspects of instruction / curriculum to be improved	-0.13	-0.24	-0.18
	Compare the school to other schools	0.09	-0.03	0.12
Does your school provide information to parents….	..**on their child's academic performance relative to other students in your school?**	**-0.36**	-0.31	**-0.36**
	..on their child's academic performance relative to national / regional benchmarks?	-0.05	-0.23	-0.11
	..on the academic performance of students as a group relative to students in the same grade in other schools?	-0.05	-0.18	-0.12
In your school are achievement data used for any of the following accountability purposes?	**Posted publicly**	**0.36**	0.17	0.28
	Evaluation of the principal's performance	-0.12	-0.23	-0.11
	Evaluation of teachers' performance	**-0.37**	**-0.40**	-0.31
	Decisions about resource allocation	0.06	-0.20	-0.09
	Tracked over time by administrative authority	-0.07	**-0.36**	-0.18

Finally, two documents produced by the European Commission gave some indicators of national assessment systems for European Union countries (EACEA, 2009a; EACEA, 2009b). These mainly related to the existence or not of particular aspects of assessment systems and their aims and uses. According to the documents the information for each country was derived from legislation, national regulation or other official documents and was provided by the national units of the Eurydice network. These units are usually based in the education ministry so the information should be reliable (see EACEA 2009b, p10, for further details on data collection).

These indicators are listed below. It should be noted that the final three variables were actually derived from combining some classifications in the original data (countries classified as either having certified assessment based on external exams or having certified assessment based on internal assessment and external exams were combined).

1) A main aim of nationally standardised tests is to take decisions about the school career of students
2) A main aim of nationally standardised tests is to monitor schools and/or the education system
3) A main aim of nationally standardised tests is to identify individual learning needs
4) Test results are used in external evaluation of schools
5) The country has recommendations or support tools for the use of test results during internal school evaluation
6) Test results are not used for external evaluation and no recommendations / support for use of test results in internal evaluation
7) Publication of individual school results in national tests is organised or required of schools by central or local government
8) Certified assessment at the end of *lower* secondary education is by final grade based on an external exam or a combination of internal assessment and external exam.
9) Certified assessment at the end of *upper* secondary education is by final grade based on an external exam or a combination of internal assessment and external exam.
10) Certified assessment at the end of *either lower or upper* secondary education is by final grade based on an external exam or a combination of internal assessment and external exam.

This data was only available for European Union countries, so models that included these variables were run using a reduced dataset. This reduced the number of countries to 24 and the number of students to 111,902.

Again, five models were run for each test, one for each plausible value and the parameter estimates from each model were combined to give an overall average parameter estimate.

Results

The following section provides some more descriptive data about the variables related to assessment systems.

PISA school questionnaire variables

Table 7 presents some descriptive statistics for the PISA questions at the country level. The data relates to the percentage of students within the country in schools with the stated policy. For each question, the figure for the UK is presented separately, along with the average, standard deviation, minimum and maximum across all countries.

For these questions the data for France was missing and data for the question about providing information relative to other students was missing for Denmark.

Table 7: Descriptive data for assessment related variables

In your school….	UK percentage	Average percentage within country	Standard deviation between countries	Minimum percentage reported by any country	Maximum percentage reported by any country
Assessments are used to monitor the school's progress from year to year.	97.0	78.8	18.4	35.6	97.9
Assessments are used to make judgments about teacher effectiveness.	82.7	49.2	23.4	9.5	92.3
Information is provided to parents about performance relative to other students.	34.6	45.3	19.3	11.2	87.9
Achievement data are posted publicly.	80.1	33.7	23.1	1.8	89.3
Achievement data are used in evaluation of teachers' performance.	94.2	46.5	27.1	0.0	87.6
Achievement data are tracked over time by an administrative authority.	93.7	67.1	19.5	29.2	95.6

Thus, the average percentage of students (across countries) in schools that use assessments to monitor progress from year to year was 78.8%. This is less than the percentage in the UK (97.0%). The most common policies in schools were using assessment data to monitor the school's progress over time and for achievement data to be tracked over time by an administrative authority.

Appendix B presents a series of plots of each country's PISA score against the percentage of students within a country studying in schools with the stated policy. This gives an indication of the relationship between school policies (at a national level) and PISA scores, but takes no account of background variables in the country. Inspection of these plots suggests that the strongest relationships (both negative) were between PISA scores and the percentage of students in schools where assessments are used to make judgments about teacher effectiveness and in schools where achievement data is used in the evaluation of teachers' performance (this is not surprising as the two variables measure very similar things).

EACEA variables

Table 8 summarises these variables, giving the mean PISA scores for each country by whether or not they have the stated policy within their country. The most notable differences were that countries where a main aim of nationally standardised tests was to take decisions about the school career of students, those where publication of individual school results in national tests was required of schools and those with an external exam at the end of lower secondary school had higher average scores across all three tests. However, none of the differences in average scores were statistically significant, as judged by a T-test.

Table 8: Mean PISA scores for each country by whether or not they have the stated policy within their country

			Mean PISA score		
		No of countries	Read	Maths	Science
A main aim of nationally standardised tests is to take decisions about the school career of students	Yes	10	497.0	499.7	506.7
	No	14	485.4	491.6	496.9
A main aim of nationally standardised tests is to monitor schools and/or the education system.	Yes	17	488.3	492.4	500.6
	No	7	494.9	501.4	501.9
A main aim of nationally standardised tests is to identify individual learning needs	Yes	8	496.9	496.0	501.6
	No	16	486.9	494.5	500.6
Test results are used in external evaluation of schools	Yes	5	495.6	497.0	504.6
	No	19	488.8	494.5	500.0
Country has recommendations or support tools for the use of test results during internal school evaluation	Yes	8	491.9	497.6	505.5
	No	16	489.4	493.7	498.7
Test results not used for external evaluation and no recommendations / support for use of test results in internal evaluation	Yes	11	488.5	494.0	499.4
	No	13	491.6	495.8	502.3
Publication of individual school results in national tests is organised or required of schools by central or local government	Yes	8	498.6	502.4	508.1
	No	16	486.0	491.3	497.4
Certified assessment at the end of lower secondary education is by final grade based on an external exam or a combination of internal assessment and external exam	Yes	11	496.9	498.7	505.5
	No	13	484.5	491.8	497.2
Certified assessment at the end of upper secondary education is by final grade based on an external exam or a combination of internal assessment and external exam	Yes	18	490.3	493.9	502.6
	No	6	489.8	498.2	496.0
Certified assessment at the end of *either lower or upper* secondary education is by final grade based on an external exam or a combination of internal assessment and external exam	Yes	19	490.8	494.6	502.3
	No	5	487.7	496.4	496.0

Modelling

For the multilevel models, each of the variables related to assessment systems was included in a separate model for each of the three tests. Additionally, for the variables taken from the PISA school questionnaire, two separate sets of models were run. The first of these was with the variable added at the school level (i.e. whether or not the school uses assessment data in the stated way). The second was with a derived country level variable, being the percentage of pupils in the country attending a school that uses assessment data in the stated way. Pupil background characteristics were included as covariates in each of the multilevel models as before. It should be noted that the data set used for the modelling was restricted to students with complete information. That is, students with missing information on any of the background variables of interest were deleted listwise. This reduced the

number of students from 235,305 to 176,886. It also meant that all data from two countries (France and Denmark) were removed.

Table 9 presents a summary of the parameter estimates for each of the assessment system variables. Asterisks indicate statistical significance at the 5% (*) and 1% (**) level.

Table 9: Parameter estimates for variables relating to assessment systems

Variable	Read	Maths	Science
Assessments used to monitor schools progress over time (% of students in country)	-0.30	-0.73	-0.41
Assessments used to make judgments about teacher effectiveness (% of students in country)	-0.48*	-0.57*	-0.45*
Info to parents on performance relative to other students in school (% of students in country)	-0.45*	-0.48	-0.52*
Achievement data posted publicly (% of students in country)	0.39*	0.21	0.34
Achievement data used to evaluate teachers (% of students in country)	-0.35*	-0.46*	-0.34
Achievement data tracked over time (% of students in country)	-0.09	-0.56*	-0.27
Assessments used to monitor schools progress over time (school level)	1.06	1.27	-0.18
Assessments used to make judgments about teacher effectiveness (school level)	1.12	-0.45	-0.37
Info to parents on performance relative to other students in school (school level)	2.90**	3.90**	3.47**
Achievement data posted publicly (school level)	8.48**	7.83**	7.52**
Achievement data used to evaluate teachers (school level)	1.99	0.74	0.59
Achievement data tracked over time (school level)	0.82	1.39	0.80
Standards-based external exams	-4.72	-5.38	-2.39
A main aim of standardised tests is to take decisions about the school career of students	16.49	9.53	14.16
A main aim of standardised tests is to monitor schools and/or the education system	-19.74*	-18.28	-14.33
A main aim of nationally standardised tests is to identify individual learning needs	9.02	-5.07	-2.16
Test results are used in external evaluation of schools	13.37	5.29	8.29
Country has recommendations or support tools for the use of test results during internal school evaluation	-2.02	-0.57	1.22
Test results not used for external evaluation / no recommendations / support for use of test results in internal evaluation	-0.32	1.66	1.90
Schools are required to publish results in national tests	13.27	8.34	9.52
Country has final external exams at the end of lower secondary education	15.48	6.31	10.83
Country has final external exams at the end of upper secondary education	-2.58	-7.95	4.66
Country has final external exams at the end of either lower or upper secondary education	-1.84	-9.27	1.34

The first result worth noting is that the models found no significant impact of the existence of standards-based external exams in the country (PISA measure), after accounting for background variables. In fact, in all three tests the parameter estimate was negative,

pointing towards lower (but not statistically significantly lower) PISA scores in countries with these exams. This finding contradicts some previous work that looked at the performance of students in countries with and without these exams (Bishop, 1997; Wößmann, 2003; Fuchs and Wößmann, 2007). However, it is consistent with the work of Mons (2009) in showing that the impact of external examinations is not clear cut. Also it is consistent with the work of Jurges and Schneider (2004) in that our analysis confirms that once we restrict analysis to a set of developed countries (in our case countries similar to the UK, in their case OECD countries only) the impact of external examinations is reduced.

In terms of the other variables taken from the PISA data, there was one with a consistent significant effect at the country level across all three tests. This was the percentage of students in schools using assessments to make judgments about teacher effectiveness, which was negatively related to PISA scores. The parameter estimates were all about -0.5, meaning that an increase in the percentage of schools using data in this way by 1% was associated with a fall in PISA scores of 0.5 of a point. This is equivalent to saying that students in a country with 75% of schools doing this would have a PISA score about 25 points lower on average than those in a country with only 25% of schools doing so. The other significant parameter estimates for the PISA variables (at country level) were also mostly negative, and of a similar magnitude. The percentage of students in schools using achievement data to evaluate teachers was negatively related to PISA scores (significantly so for reading and maths). Similarly, the percentage of students in schools providing information to parents about performance of their children relative to other students was negatively related to PISA scores (significant for reading and science).

The only significantly positive country level coefficient is for the percentage of students in schools posting achievement data publicly, which was significant for reading only. A 1% increase in this improved the PISA reading score by 0.39 points on average.

It is interesting that when looking at the effects of the PISA variables at school level a slightly different picture emerges, with the only significant parameter estimates being positive. For instance, students in schools which provide information to parents about the performance of their children relative to other students had significantly higher PISA scores on average (by about 3-4 points) after accounting for background variables. This contrasts with the negative coefficient for this variable at country level. It is not clear why this apparent contradiction occurs, but it may be that, within a country, schools that provide this information do so because they have more students with parents who are competitive or ‘pushy’ and demand it. This could mean that schools with high attaining students would be more likely to be required to provide such information. In other words, the high attainment of students may be driving the need to provide data to parents rather than the provision of data having a positive impact.

The other significant positive effect at the school level was posting achievement data publicly, which was associated with an increase in PISA scores of around 7-8 points in each test. This supports the findings for this variable at the country level (although this effect was small and only significant for reading).

Of the variables taken from the EACEA reports only one had a significant effect, and this was for the reading test only. Students in countries where a main aim of standardised tests is to monitor schools and/or the education system tended to perform worse than students in countries without that aim (or with no standardised tests). This effect was worth around 20

points in reading, 18 in maths (not significant) and 14 in science (n.s.). Note that this question is not asking whether standardised tests are *used* in monitoring and evaluation but whether this is a main aim of the assessments. Certain countries (such as the Netherlands) make use of national test data in school monitoring but maintain that the *main* aim of such assessments is to make decisions about the school careers of pupils.

It is also interesting to note that students in countries with final external exams at the end of lower secondary education tended to perform better (by about 15 points in reading, 6 points in maths and 11 points in science, all not statistically significant). This contrasts to some degree with the parameter estimates for the PISA measure of the existence of standards-based external exit exams (all negative, although also not significant). It is not clear why this apparent contradiction occurs, but it could be due to the different countries included in the models. To test for this, the model including the PISA measure variable was re-run, but for European countries only. The new parameter estimates were 0.00 for reading, -0.89 for maths and 3.47 for science. Thus, removing non-European countries meant that the negative association between external exams and PISA scores almost completely disappeared. However, it did not become a positive association which would be consistent with the results from the EACEA variable. This might be due to the variable taken from the EACEA report referring to lower-secondary education only. Seven of the European countries did not have external exit exams according to PISA, compared with 11 countries not having these exams at lower secondary according to EACEA. The difference between the two sets of results may also simply reflect the small numbers of countries available for analysis and the level of uncertainty surrounding each of our estimates.

The other variables taken from the EACEA reports all had positive association with PISA scores, but none of the coefficients achieved statistical significance. However, it is worth noting that students in countries where schools were required to publish results in national tests tended to perform better (although not statistically significant). This is consistent with the positive effect at school level of publishing achievement data publicly. Again, it should be noted that the EACEA variables were only available for EU countries and they may have had a different effect if data for other countries had been used.

Some further analyses were undertaken after removing data from three countries which were clear outliers in terms of their mean PISA scores. These can be seen in Appendix B, in the graphs comparing PISA scores with the percentage of students in each country in schools with particular policies. The three countries with the lowest PISA scores often do not fit in with the pattern of data points. The countries were identified as Miranda-Venezuela, Trinidad and Tobago and Bulgaria. To see what impact these countries were having on the overall models it was decided to remove them from the data and re-run the models. Table 10 presents the parameter estimates for the assessment system variables for each model with the reduced data. This analysis may in fact provide a more robust assessment of the likely impact of assessment systems within countries such as the UK.

The estimated school level coefficients within Table 10 are very similar to those in the original analysis in Table 9. However, at country level some of the changes are worth mentioning. In general the effect was to reduce the size of the parameter estimates by a small amount. The percentage of students in schools using assessments to make judgments about teacher effectiveness was no longer a significant factor, and the percentage of students in schools providing information to parents about performance of their children

relative to other students was no longer significant for any of the tests (and was reduced almost to zero). This means that the contradiction commented on above in relation to providing information to parents (negatively related at country level, positively related at school level) was no longer present. This factor was now positively related to PISA scores only, when considered at the school level. Finally, the one EACEA variable with a significant parameter estimate in Table 9 ('a main aim of standardised tests is to monitor schools and/or the education system') was non-significant in the models excluding outlying countries.

Table 10: Parameter estimates for variables relating to assessment systems (outlying countries removed)

Variable	Read	Maths	Science
Assessments used to monitor schools progress over time (% of students in country)	-0.13	-0.55*	-0.22
Assessments used to make judgments about teacher effectiveness (% of students in country)	-0.24	-0.27	-0.16
Info to parents on performance relative to other students in school (% of students in country)	-0.02	0.00	-0.04
Achievement data posted publicly (% of students in country)	0.23	0.02	0.13
Achievement data used to evaluate teachers (% of students in country)	-0.20	-0.32*	-0.18
Achievement data tracked over time (% of students in country)	0.06	-0.38	-0.11
Assessments used to monitor schools progress over time (school level)	0.74	0.80	-0.85
Assessments used to make judgments about teacher effectiveness (school level)	0.48	-0.87	-0.98
Info to parents on performance relative to other students in school (school level)	2.44*	3.50**	3.06**
Achievement data posted publicly (school level)	7.51**	6.94**	6.37**
Achievement data used to evaluate teachers (school level)	1.24	0.30	-0.05
Achievement data tracked over time (school level)	0.58	1.17	0.64
Standards-based external exams	1.70	1.29	4.94
A main aim of standardised tests is to take decisions about the school career of students	13.22	5.66	10.71
A main aim of standardised tests is to monitor schools and/or the education system	-16.97	-15.20	-11.41
A main aim of nationally standardised tests is to identify individual learning needs	6.02	-8.58	-5.35
Test results are used in external evaluation of schools	10.68	2.21	5.49
Country has recommendations or support tools for the use of test results during internal school evaluation	-5.44	-4.32	-2.21
Test results not used for external evaluation / no recommendations / support for use of test results in internal evaluation	5.17	7.91	7.73
Schools are required to publish results in national tests	10.38	5.02	6.45
Country has final external exams at the end of lower secondary education	12.25	2.41	7.37
Country has final external exams at the end of upper secondary education	0.43	-4.75	7.84
Country has final external exams at the end of either lower or upper secondary education	1.02	-6.23	4.31

Conclusions

Using data from international studies to explore the potential for elements of an education system to provide improvement in a UK context is not straightforward. To begin with, once we analyse the background characteristics of students, we find that for many countries the characteristics of students are too different from those of their UK counterparts for any comparisons to be made. If anything, our analysis in this report has shown that other countries are even less "comparable" to the UK than might be initially expected. This is an important fact to bear in mind if we are to make inferences from international data about which changes to our education system might be most beneficial. By restricting our attention to those countries most similar to the UK we improve our chances of making conclusions from the data that are relevant within our own context.

Furthermore, we have seen that the international rankings of countries are to some extent driven by the background characteristics of students. Once these background factors are accounted for in analysis there can be substantial changes in the estimated performance of countries relative to one another; in particular the relative performance of the UK substantially improves. This implies that adjusting for the impact of student characteristics is of crucial importance if we are to discover which elements of an assessment system are truly influential.

The main motivation for our research was to explore the impact of various facets of an assessment system. Our analysis shows that the evidence for the capacity of assessment systems to become a major driver of improvement in educational systems is hardly overwhelming. At the country level, once outliers are removed, only two statistically significant effects were identified relating the details of countries' assessment system to their performance in PISA 2009 (both in maths only). Indeed, once we restrict ourselves to countries with students' similar to those in the UK, there is little convincing evidence that the existence of standards-based external exams has a positive impact on the achievement of pupils at all. Among countries with standards-based external exams we can find high performers (such as the Netherlands) and low performers (such as Israel). Similarly, amongst those countries without standards-based external exams we can find high performers (such as Belgium) and low performers (such as Greece). Often within policy debates (perhaps out of politeness) the education systems of low performing countries are ignored. However, it is only when we take account of all of the data (including the low performers) and, furthermore, attempt to adjust for other influential factors that we can get a true picture of the influence of a system level variable; albeit an inconclusive one.

Both the statistically significant effects found within our multilevel models were to do with monitoring and evaluation; and both times the effect was found to be negative. That is, particularly in mathematics, students in countries where the use of assessment data to evaluate teacher and school performance is widespread tend to perform slightly less well than similar students elsewhere. The reasons for these negative coefficients are unclear. A possible hypothesis might be that if teachers perceive the main purpose of assessments as being to evaluate their own performance this may increase their desire to "teach to the test" and thus harm the extent to which their students receive a uniform and broad education in their given subject. In mathematics in particular it might be that an over emphasis on teacher evaluation may lead to teachers drilling their students to be able to perform particular tasks that they suspect will be assessed. This may lead to less emphasis on providing the broad mathematical problem solving skills required to perform effectively in questions on the

practical application of mathematics such as those typical of PISA assessments. Results based on EACEA variables within European countries provide a subtly different emphasis. It is only when the *main* aim of national assessments becomes to monitor schools and/or the education system that it is found to have a negative effect on performance (although not statistically significant). Overall, countries that use test results in external evaluation of schools, whether or not this is a main aim, tend to outperform countries that do not (again, not statistically significantly). This would imply that using assessment results for evaluation is acceptable (and potentially even beneficial) provided that the assessments themselves have the interests of learners at their centre and will not create perverse incentives for teachers to behave in a pedagogically unsound manner.

In the light of the above paragraph, it is interesting to note that despite the negative effects noted above, the widespread publication of school level results was not found to be negatively associated with performance. This implies that the negative effects of an over-emphasis on exam performance for teacher and school evaluation are to some extent ameliorated when performance data is public. On the one hand this may indicate that the competition between schools that is created by publicly available achievement data serves to drive up standards and thus counteracts the possible negative effects of an overemphasis on accountability. Even if this were true, it would not imply that competition actually improves results, only that we avoid the negative effects of assessment for teacher evaluation. This would imply that, having created all the machinery of collating and publishing schools' results and all of the additional stress for the school workforce that goes with it, the best we could say for certain is that overall we have probably not damaged student ability in reading, maths and science. An alternative explanation, restricting our attention to the European data, would be that it is possible that monitoring and evaluation, potentially including publishing individual school results, may in fact be beneficial provided that this is not the main aim of the assessments.

Limitations

Finally, it is worth summarising the limitations and assumptions made in this research. Firstly, across all the methods used in the analyses the assumption was that the background variables we controlled for have a similar meaning across countries. So, for example, having three cars in your home has the same sort of implications in Australia in terms of the socio-economic background of different students as it does in the UK. This point applies to the method used to select countries similar to the UK as well to the multilevel modelling, since these background characteristics are controlled for in both analyses.

In selecting countries similar to the UK there was a trade-off between choosing enough countries for a robust statistical analysis and choosing those most similar. The results of the effective sample size method demonstrated that there were no countries which provided unproblematic comparisons with the UK. This meant that in order to select a reasonable number of countries we included some which were on the face of it not very similar. This is certainly something that should be taken into account when using international data to make comparisons, even to supposedly 'similar' countries.

A further, more general, note of caution should be noted in interpreting the data. Whilst there is evidence of an association between some aspects of assessment systems and performance on PISA, there is no evidence that the relationship is causal, or of the direction of causality. It might be that, for instance, countries performing better on PISA are more

likely to publish achievement data publicly. Or it could be that both PISA scores and publishing achievement data are related to a different, unknown variable.

It is also acknowledged that it is possible the results presented here were dependent on the exact methodology being used, and using an alternative methodology would generate substantially different results and conclusions. In order to investigate this, a number of sensitivity analyses were undertaken, using different methods to analyse the data. The results of these analyses are presented in Appendix C. In general they show that using different methods would have had only minimal impact on the results and so we can be fairly confident that the conclusions drawn here are valid.

References

Bishop, J. H. (1997). *The Effect of National Standard and Curriculum-Based Exams on Achievement*. CAHRS Working Paper Series.

Crisp, V. (2013). *Cultural and societal factors in high-performing jurisdictions*. Cambridge Assessment research report. Cambridge: Cambridge Assessment.

EACEA (2009a). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Brussels: Eurydice.

EACEA (2009b). *Key Data on Education in Europe 2009*. Brussels: Eurydice.

Fuchs, T., and Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics,* 32, 433–464.

Hanushek, E. A. and Wößmann, L. (2011). The Economics of International Differences in Educational Achievement. In E.A. Hanushek, S. Machin and L. Wößmann (Eds.) *Handbook of the Economics of Education, Volume 3.* San Diego, CA. Elsevier.

Jurges, H. and Schneider, K. (2004) International Differences in Student Achievement: An Economic Perspective. *German Economic Review*, 5, 3, 357–380.

Jurges, H., Schneider, K. and Buchel, F. (2003). *The effect of central exit exams on student achievement: Quasi-experimental evidence from TIMSS Germany*. CESifo working paper No. 939. CES: Munich.

Mons, N. (2009). *Theoretical and real effects of standardised assessment*. Brussels: Eurydice.

OECD (2007) PISA 2006: *Science Competencies for Tomorrow's World*. OECD: Paris.

OECD (2009). *PISA Data Analysis Manual: SAS Second Edition.* OECD: Paris.

OECD (2010). *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practice.* OECD: Paris.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics,* 65, 117–170.

Wößmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, 13, 2, 143-169.

Appendix A: Technical formulae for calculating effective sample sizes

The aim of the effective sample sizes calculation is to estimate the extent to which applying the weights required to make the background characteristics of students in other countries match up with those in the UK is equivalent to reducing the size of the data. In other words, how much is the data from any particular country "worth" once it has been weighted to adjust for the background characteristics of students.

To begin with we suppose that a quantity of interest (X) is measured by responses from *n* independent students $x_1, x_2, \ldots, x_n$. Suppose further that each observation is weighted by weights $w_1, w_2, \ldots, w_n$. If we now suppose that we are interested in estimating the mean of this quantity then we know that the precision of our estimate will be determined by the variance of the estimate. If our observations are independent then this is equivalent to:

$$V(Estimate) = V\left(\frac{\sum w_i x_i}{\sum w_i}\right) = \frac{\sum w_i^2 V(x_i)}{(\sum w_i)^2} = V(x_i)\frac{\sum w_i^2}{(\sum w_i)^2}$$

Next we note that for a simple random sample the variance of an estimate of the mean would be:

$$V(Estimate) = V(x_i)\frac{1}{n}$$

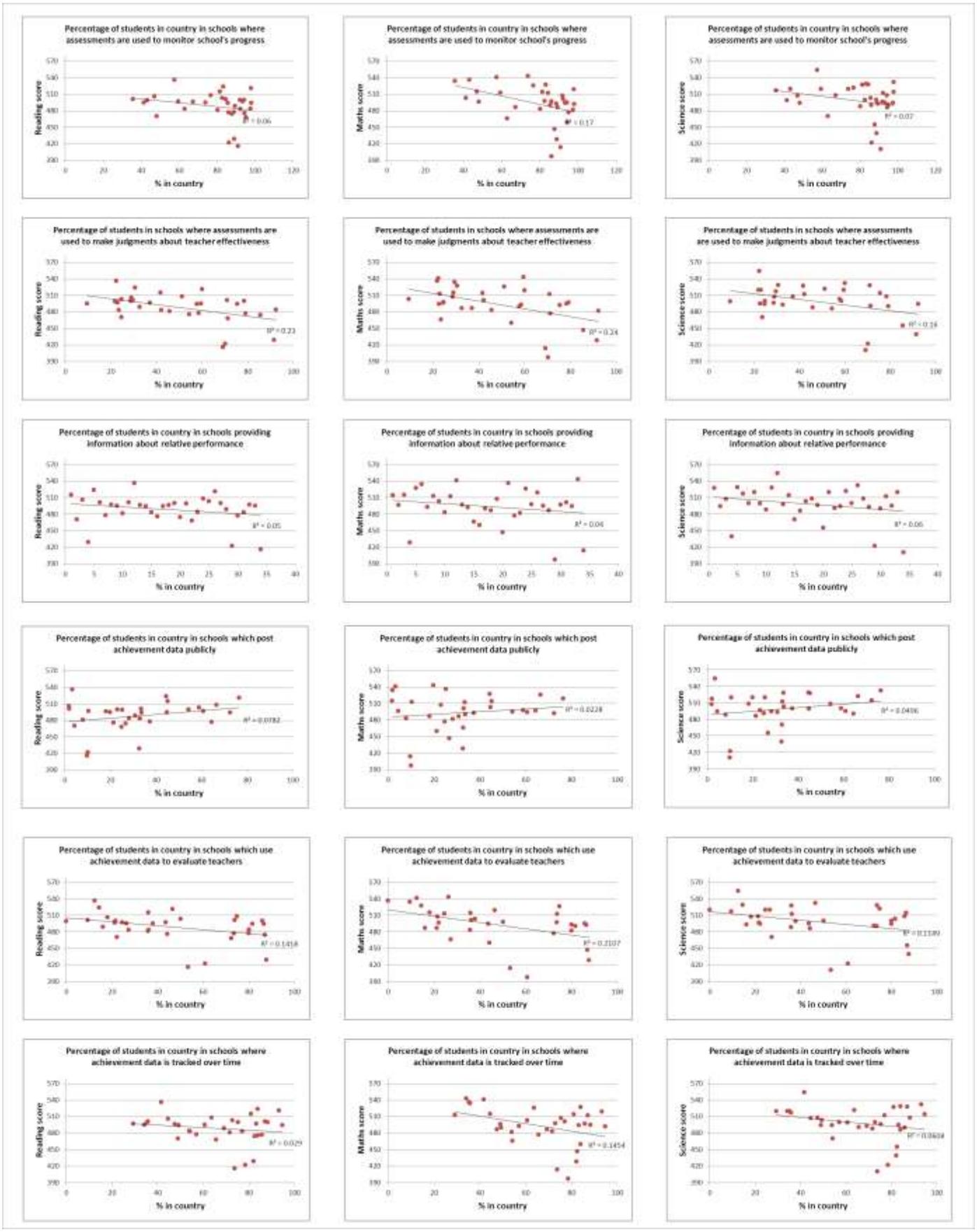
In other words, the variance of a weighted sample is equivalent to the variance of a simple random sample of size:

$$Effective\ Sample\ Size = \frac{(\sum w_i)^2}{\sum w_i^2}$$

This quantity will be largest when all weights are equal (that is, our sample is unweighted) in which case the effective sample size will simply equal *n*; the actual sample size. Thus, this formula estimates the impact of any weighting upon the accuracy of estimates from a simple random sample. This is quantified in terms of the equivalent reduction in sample size.

It should be noted that in the PISA data set observations are not independent of one another; in particular there are likely to be correlations between the responses of pupils in the same schools. This means that the above formula does not provide an exact estimate of the impact of weighting on the precision of estimates from different countries. However, the above formula still provides a reasonable formula for exploring the extent of differences between the UK and other countries. If there was a country where students' characteristics were already similar to those of students in the UK, this country's data would not require weighting and so the effective sample size would be equal to the original sample size. However, countries that require extreme weighting in order to be made comparable to the UK will have low effective sample sizes.

Appendix B: PISA score against the percentage of students within a country studying in schools with the stated policy



Appendix C: Sensitivity analyses

A number of analyses were conducted to verify that the conclusions presented within this research are not overly dependent upon the exact methodology used within this report. Where alternative methods have been considered the impact of the decision to use one method rather than another is considered below.

Use of weighted data

All of the main multilevel models described previously were based on unweighted data. Unweighted data was used to ensure that the standard errors of model coefficients were kept as low as possible and thus the power of analysis was maximised. However, one potential drawback of this approach is that it means the data used in analysis may be slightly less representative of students within any particular country than if weights were used. Two pieces of sensitivity analysis were undertaken to determine the impact of this decision upon results.

Matching alternative countries to the UK using weighted UK data rather than unweighted UK respondents

A first area for investigation was to explore how the selection of countries to include within the analysis would change if we were to match the characteristics of students from other countries to the characteristics of UK students within the weighted rather than unweighted data. To explore the potential impact of this change propensity scores were recalculated for students in each country based on matching characteristics to the weighted UK data. To further explore the sensitivity of results to the statistical methodology that was used, weights were calculated within this method by sub-classifying students into one of 15 groups on the basis of their propensity score and then calculating weights on the basis of the proportions of students in each group[6] (this method of weighting is recommended by Rosenbaum and Rubin, 1984). Effective sample sizes for each country were then recalculated on the basis of these revised propensity scores. Across all countries there was found to be a correlation of 0.9 between the originally calculated effective sample sizes and the effective sample sizes calculated by this revised method. This indicates that the selection of countries to include in analysis is largely robust to the method used to select them.

Weighted Multilevel Modelling

As described earlier, all of the multilevel modelling was applied to unweighted data to ensure that the standard errors of analysis were as low as possible. However, one effect of this is that, when estimating the effect of individual covariates in order to adjust the overall country scores for their effect, countries with a greater number of respondents will have a greater influence over the size of coefficients than smaller countries. Furthermore, the students within each country used to estimate coefficients may not be representative of students in their country as a whole. To examine the impact of using unweighted data the initial multilevel models (those that were used to produce a revised country ranking) were rerun using the original weights supplied within the PISA dataset to weight the analysis and also giving equal weight to students in each country. For each of the outcome scores (Reading, Maths and Science) the correlation between the original country effects and the country effects based on weighted multilevel modelling was calculated. The correlation was found to be equal to 0.997 in each case indicating that our results are robust to the decision as to whether weighted or unweighted data should be used.

Adjusting for impact of contextual variables using propensity score matching rather than multilevel modelling

One potential criticism of using multilevel modelling to adjust for the background characteristics of students in each country is that it assumes a fairly uncomplicated relationship between each of the covariates and the outcome of interest. For example, our model does not account for interactions

[6] This is in contrast to the inverse probability weighting used in the original analysis.

between the different background characteristics or take account of potential nonlinear effects. To address this criticism a potential alternative method would be simply to use the weights generated via the propensity scores to make the background characteristics of each country equivalent to begin with. Once this is done we can simply compare the mean scores between countries knowing that we are making comparisons on a like-with-like basis without the need to assume a particular functional form for the relationship between background variables and the outcomes of interest[7]. The drawback with this approach is that the propensity score matching only takes account of 10 of the background variables.

The country effect coefficients derived from the original multilevel modelling were compared to average country scores in each subject derived by each of two propensity score matching methods: matching to unweighted UK data (original propensity score weights) and matching to weighted UK data.

Original weights from propensity score matching

The country effect coefficients derived from the original multilevel modelling were compared to average country scores in each subject weighted using the original propensity scores; that is, those created via matching to unweighted UK data. Generally there was a reasonably high level of agreement between these measures with correlations of between 0.91 and 0.95 being found between contextually adjusted PISA scores using weights and the overall country level effects derived via MLM. However, results for two countries (Trinidad and Tobago and Liechtenstein) appeared particularly sensitive to the choice of method. In the latter case (Liechtenstein) this is probably a result of the small overall sample size available from this country. The reasons for the sensitivity of results from Trinidad and Tobago to the choice of method were not clear. Once these two countries were removed, the correlations between country effects based on multilevel modelling and weighted average scores were 0.93, 0.97 and 0.94 for Reading, Maths and Science respectively. Furthermore, these correlations were higher than the correlations between the original PISA scores and the contextually adjusted scores derived by either method, indicating that both methods of contextual adjustment are to some extent performing the same task. Nonetheless, it could be seen that the choice of method used to adjust for student context does have some impact on results. The results from the two methods are not identical, and this fact becomes all the more true if we report results in terms of ranks rather than as scores. This implies that, if we wish to adjust for the impact of student background, then the choice of which variables we adjust for and the method we use will affect our results. Ultimately we have used multilevel modelling as our main method within analysis as it allowed us to adjust for a greater number of variables. However, we need to be aware that our results (particularly in terms of contextually adjusted country rankings) are somewhat dependent upon the method we have chosen. Whilst we can confidently conclude that taking account of the background of students can lead to important changes in the relative positions of countries, we cannot necessarily conclude that our own ranking is more valid than any other that might be derived via different methods.

Weights based on grouped propensity scores to match to weighted UK data

For each of the three subjects (Reading, Maths and Science) the contextually adjusted country level scores based on this method of propensity score matching had a correlation of more than 0.99 with the contextually adjusted country level scores derived via the previous method of propensity score matching. This strongly indicates that the decisions to match to unweighted rather than weighted UK data, and to directly apply propensity scores rather than match countries on propensity score groups, has no significant impact on the results.

[7] Although we do need to implicitly assume a particular functional form for the relationship between the background characteristics and whether or not students are from the UK.

Listwise deletion of candidates with missing background information

For the purposes of many of our analyses we have restricted our attention to students with complete information. That is, students with missing information on any of our background variables of interest were deleted listwise. In order to check the potential impact of this restriction on analysis, country mean scores based on students with listwise complete background data were compared to the original mean scores published for each country within the main PISA report. For the set of 35 countries included within our analysis, correlations of between 0.97, 0.99 and 0.98 were found for Reading, Maths and Science respectively[8]. This indicates that the decision to limit analysis to pupils with listwise complete background data does not have a significant impact on our results.

Reference

Rosenbaum, P. and Rubin, D. (1984) Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79 (387), 516-524.

[8] Although, interestingly, across all countries pupils with listwise complete data were found to have higher scores on average than those with missing data.