Department of Quantitative Social Science

England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline?

John Jerrim

DoQSS Working Paper No. 11-09
December 2011

Leading education and social research
Institute of Education
University of London

DISCLAIMER

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science. Institute of Education, University of London. 20 Bedford way, London WC1H 0AL, UK.

England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline?

John Jerrim[*†]

Abstract. The Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) are two highly respected cross-national studies of pupil achievement. These have been specifically designed to study how different countries' educational systems are performing against one another, and how this is changing over time. These are, however, politically sensitive issues, where different surveys can produce markedly different results. This is shown via a case study for England, where apparent decline in PISA test performance has caused policymakers much concern. Results suggest that England's drop in the PISA international ranking is not replicated in TIMSS, and that this contrast may well be due to data limitations in both surveys. Consequently, I argue that the current coalition government should not base educational policies on the assumption that the performance of England's secondary school pupils has declined (relative to that of its international competitors) over the past decade.

JEL classification: I20, I21, I28.

Keywords: PISA, TIMSS, educational policy, change over time.

[*]Institute of Education, University of London. E-mail: J.Jerrim@ioe.ac.uk.

[†]I would like to thank Anna Vignoles and John Micklewright for their helpful comments on earlier versions of this work. Feedback has also been gratefully received from Jenny Bradshaw (NFER), Emily Knowles, Helen Evans and Lorna Bertrand (Department of Education). The views in this paper do not represent those of the Institute of Education and are the author's alone.

1. Introduction

One of the major developments in educational research over recent years has been the widespread implementation of the international studies of pupil achievement, PISA, TIMSS and PIRLS[1]. Each has the explicit aim of producing cross-nationally comparable information on children's abilities at a particular age in at least one of three areas (reading, maths and science) and is widely cited by academics and policymakers. Another goal of these studies is to monitor how countries are performing relative to one another (in terms of educational achievement of school pupils) over time. An example is a recent report published by the Organisation for Economic Co-operation and Development (OECD 2010a), which used information from the four waves of PISA to investigate how test scores have changed across countries since 2000. In this paper I provide a similar case study for one country, England, where the issue of change in performance in the international achievement tests has had a large impact on government officials' thinking and public policy debate.

It is important to explain my motivation for focusing on England and why this has become such an important (and politically sensitive) issue. Children who took part in the first PISA wave (2000) were born in 1984, and would thus have received most of their compulsory schooling during the years when the Conservative party was in power (who held office between 1979 and 1997). The majority of the most recent (PISA 2009) cohort were, on the other hand, born in 1994, and so spent all their time in school under Labour (who governed between 1997 and 2010). Whether rightly or wrongly, many commentators have thus regarded change in England's PISA ranking since 2000 as an evaluation of the Labour government's educational policy success.

When the PISA 2009 results were released in December 2010, it was therefore England's dramatic decline in performance that grabbed the domestic headlines. Figure 1 highlights exactly why this happened. The solid grey line refers to change in real educational expenditure in England since 2000 with dashed lines referring to mean PISA maths test

[1] PISA stands for the Programme for International Student Assessment, TIMSS for Trends in International Mathematics and Science Study and PIRLS for the Progress in International Reading Literacy Study. The former is co-ordinated by the OECD and the latter two by the International Association for the Evaluation of Educational Achievement (IEA). This paper focuses on PISA and TIMSS.

scores (author's calculations) over the same period[2]. A dotted line is also included to illustrate the change in the proportion of children who obtained 5 or more A*-C grades in their GCSE exams (or equivalents)[3]. Figures refer to index values, with 2000 set as the base year and assigned a value of 100. As one can see, spending on education rose by around 25% over this period in real terms, and was accompanied by a large increase in the proportion of young people achieving 5 A*-C grades. Yet the PISA data contradict this pattern, suggesting that England's secondary school pupils' average maths performance has been in relative decline.

Figure 1

This has since become a widely cited "fact" that has been used for both political benefit and to justify the need for policy change. The Daily Telegraph (a leading English newspaper) ran a commentary stating that[4]:

"This is conclusive proof that Labour's claim to have improved Britain's schools during its period in office is utter nonsense. Spending on education increased by £30 billion under the last government, yet between 2000-09 British schoolchildren plummeted in the international league tables."

A sentiment echoed by the Secretary of State for Education (Michael Gove MP) in a recent parliamentary debate[5]:

"The PISA figures … show that … the standard of education … offered to young people in this country declined relative to our international competitors. Literacy, down; numeracy, down; science, down: fail, fail, fail."

The Prime Minister (David Cameron MP) and his deputy (Nick Clegg MP) have also pointed to this fall in the international rankings as one of the main reasons why England's schooling

[2] Whereas the national exam scores plotted in Figure 1 are collected annually, the PISA trend is based on data collected every three years (starting from 2000).
[3] This stands for the General Certificate of Secondary Education, important national exams children in England sit at age 15/16. The 5 A*-C threshold is often referred to by policymakers and is often treated as the minimum target that children should attempt to meet.
[4] http://blogs.telegraph.co.uk/news/tobyyoung/100067092/british-schoolchildren-now-ranked-23rd-in-the-world-down-from-12th-in-2000/
[5] www.theyworkforyou.com/debates/?id=2011-02-07b.10.3&s=pisa+%2B+gove#g11.4

system is in desperate need of change. For instance, they opened the 2010 Schools White Paper by stating that[6]:

"The truth is, at the moment we are standing still while others race past. In the most recent OECD PISA survey in 2006 we fell from 4th in the world in the 2000 survey to 14th in science, 7th to 17th in literacy, and 8th to 24th in mathematics."

It would thus seem that the change in England's PISA test scores has had an apparent influence of the formulation of the coalition government's educational policy.

But is it really true that the achievement of secondary school children in England has declined so rapidly over time (relative to other countries)? As noted by Brown et al (2007) PISA is just one study, which has its merits, but also its defects. Do other international studies of secondary school children (such as TIMSS) paint a similarly depressing picture of England's lack of progress? And if not, can the difference in results be explained?

This paper considers the robustness of the finding that secondary school children in England are rapidly losing ground relative to those in other countries. The analysis demonstrates that results from PISA and TIMSS do indeed conflict, with the latter suggesting that test scores in England have actually improved over roughly the same period. Yet the fact that these two surveys disagree with regard to change over time does not seem to be an experience that is shared (at least not to the same extent) by other countries. It is then shown how this may be due to difficulties with the PISA and TIMSS data for England, with a focus on issues such as alterations to the target population, survey procedures and problems with non-response. This leads to the following conclusions:

- Both PISA and TIMSS are problematic for studying change in average test performance in England over time.
- Statements such as those made by the policymakers cited above are based upon flawed interpretations of the underlying data.
- England's movement in the international achievement league tables neither supports nor refutes policymakers' calls for change.

[6] See http://www.education.gov.uk/schools/teachingandlearning/qualifications/gcses/b0068570/the-importance-of-teaching/pm-foreword . At the time they were writing, the PISA 2009 results had not yet been released (hence their reference to the 2006 wave).

The paper now proceeds as follows. In section 2 the PISA and TIMSS datasets are described. Section 3 provides estimates of change in test scores for England over the last decade. This is followed in section 4 by an explanation of the statistical limitations on which such estimates are based[7]. Conclusions are then presented in section 5.

2. Data

Data are drawn from the Programme for International Student Assessment (PISA) and Trends in Mathematics and Science Study (TIMSS). Both have been explicitly designed to collect comparable information on children's cognitive skills across countries and over time. The former is conducted by the OECD and examines 15 year olds in three subject areas (reading, maths and science). The latter, on the other hand, is run by the International Association for the Evaluation of Educational Achievement (IEA), with children from two different school "grades" (grades 4 and 8) being tested in science and maths. This paper focuses on the TIMSS data for the 8th grade (13 / 14 year olds – who, in reference to the English school system, take the TIMSS test towards the end of year 9).

Both studies have reasonably similar sample designs. Schools are firstly selected as the primary sampling unit and then children within these schools are chosen to participate. In TIMSS one or two classes are randomly selected, with all pupils within this class being tested. PISA, on the other hand, randomly draws 35 pupils from within each of the sampled schools. To limit non-response, both PISA and TIMSS employ a strategy of using "replacement schools". That is, if a school declines to take part, a replacement that is similar in terms of observable characteristics is asked to take its place (there is some controversy over this in the survey methodology literature - see Sturgis et al 2006 for further details). Survey weights are also produced in both PISA and TIMSS which attempt to correct for problems of non-response, while also scaling the sample up to the size of the national population. These weights are applied throughout the analysis.

Although the two studies overlap in terms of broad subject areas, there are conceptual differences in the skills they attempt to measure. Whereas TIMSS focuses on children's ability to meet an internationally agreed curriculum, PISA examines functional

[7] This paper shall focus upon three particular issues – the target population, survey procedures and non-response. Other authors have raised important issues with regard to the limitations of the international achievement datasets (e.g. Goldstein 2004) which are not covered here.

ability – how well young people can use the skills in "real life" situations. The format of the test items also varies, including the extent to which they rely on questions that are "multiple choice". Yet despite these differences, the two surveys summarise children's achievement in similar ways. Specifically, five "plausible values" are created for each child in each subject area[8]. The intuition is that children's true ability cannot be observed, and must be estimated from their answers on the test. This is done via an item-response model, although the studies do differ in their specific application of this technique (PISA uses a one parameter model while TIMSS uses a three parameter model). Brown et al (2007) provide further discussion[9]. This results in a measure of children's achievement that (in both studies) has a mean of 500 and standard deviation of 100. However, even though the two surveys appear (at face value) to share the same scale, figures are not directly comparable (e.g. a mean score of 500 in PISA is not the same as a mean score of 500 in TIMSS). This is because the two surveys contain a different pool of countries upon which these achievement scores are based (in PISA the scale is calculated in reference to OECD members while TIMSS includes a number of less developed countries). Hence one is not able to directly compare results in these two surveys (and change over time) by simply looking at the raw scores. A method for overcoming this problem is described at the end of this section.

Before doing so, I turn to some of the more specific details regarding the two surveys. The PISA study has been conducted four times (2000, 2003, 2006 and 2009) with all OECD countries taking part in every survey wave. The total number of countries in PISA has, however, risen from 43 in 2000 to 65 in 2009 (a large number of non-OECD members have been added). The implication of this is that one of the reasons England has "plummeted" down the international rankings is because more countries are now included (i.e. it is easier to come tenth in a league of 43 than it is in a league of 65)[10]. Although children were

[8] The first plausible value is used throughout the course of the analysis. Substantive findings remaining intact when other plausible values are used instead. See page 129 of OECD (2009) for further information on using just one plausible in analysis of the PISA data.

[9] Brown et al (2007) discuss the differences between the PISA and TIMSS datasets (including the response models they use) and provide greater depth to the issues overviewed here.

[10] This is only true if the additional 22 countries actually increase competition towards the upper end of the international league table. It does seem that some high-performing economies have been added (e.g. Singapore, Liechtenstein and Shanghai-China) which has pushed England's position down the overall league table. However, most of the additional countries that have been added have been those with lower levels of economic development, who come below England in the international ranking. It is also worth noting that England's performance has declined even relative to other members of the OECD.

assessed in three areas in each PISA wave, one of these was the main focus every time the survey was conducted (the so called "major domain"). In 2000 and 2009 this was reading, in 2003 maths and 2006 science. So, for instance, the inaugural study in 2000 contained around 140 items measuring children's reading skills (major domain) compared to only around 35 in each of science and maths (minor domains).

The TIMSS 8th grade study has been conducted four times (1995, 1999, 2003 and 2007), with mathematics and science skills examined using approximately the same number of questions (there is, in other words, no issue of "minor" and "major" domains). In contrast with PISA, not all of the OECD countries take part. In fact, one of the difficulties with the TIMSS data for my purposes is that a number of countries have chosen to take part in only specific survey years (e.g. data may be available in 2007, but not in, say, 1999), limiting the pool that have the relevant information available. Focus is therefore restricted to ten countries that have taken part in each of the three TIMSS (1999, 2003 and 2007) and four PISA (2000, 2003, 2006 and 2009) studies conducted since 1999. This includes four from the rich western world (Australia, England, Italy, US), a number of Asian "tiger" economies in whom policymakers have shown particular interest (Hong Kong, Japan, South Korea), and three with lower levels of development (Hungary, Indonesia, Russia). Appendix 1 loosens this strict inclusion criteria and adds six further countries into the analysis, including two from Scandinavia (Norway and Sweden), three from Europe (Czech Republic, Netherlands, Scotland) and one more rich industrialised nation from the southern hemisphere (New Zealand). The general conclusions remain largely unchanged.

Next I turn to the issue of comparability of test measures over time. Although this is a central aim of PISA, some technical details do not make this as straightforward as it first seems. In particular, the scales were only fully developed the first time a subject became a "major domain". The survey organisers therefore advise that only reading scores are actually fully comparable across all four waves, with maths becoming fully comparable from 2003 and science from 2006. See OECD (2010a page 26) for further details[11]. As can be seen from the quotes presented in the previous section, however, it is clear that this is not always

[11] The OECD has also produced a technical note on this issue which can be found at http://www.oecd.org/dataoecd/19/34/39712449.pdf

how the data are being used. At least in the case of England, policymakers always discuss change relative to performance in 2000 for all the PISA subjects.

Unfortunately, reading is not examined as part of the TIMSS study. One can therefore only compare PISA and TIMSS using either science or maths. This paper focuses on the latter as the PISA data for this subject are technically comparable over a longer period of time. Appendix 2 demonstrates that results are robust to this choice, with conclusions largely unchanged if a different PISA or TIMSS subject area or base year (i.e. if change is measured from 2003 instead of 2000) is used instead[12].

Finally, I return to the fact that PISA and TIMSS are based on a different selection of countries, meaning their test scores are not directly comparable. To overcome this problem, all data are transformed (within each survey and each wave) into international z-scores. That is, each country's mean test score (for each wave of the survey) is adjusted by subtracting the mean score achieved amongst all children in the ten countries for that particular year and dividing by the standard deviation. This is a standard method for obtaining comparable units of measurement for variables that are on different scales and was the approach taken by Brown et al (2007) in their comparison of the PISA and TIMSS datasets[13]. One implication of this is that estimates refer to English pupils' test performance relative to that of children in the other nine countries. Terms like "relative decline" shall therefore be used as international z-scores are comparative measures.

3. The change in England's maths test performance since the turn of the 21st century

The focus of this section is the change in England's maths test performance over the past decade[14]. Yet it is important to first of all consider the cross-sectional picture from TIMSS 2007 and PISA 2009. In other words, do these studies agree on how England's mean test performance currently compares? Estimates are presented in terms of international z-scores and can be found in Figure 2. The x-axis refers to PISA 2009 and the y-axis TIMSS 2007.

[12] Through the main body of this paper I compare the change between PISA 2000 and 2009 with the change in TIMSS between 1999 and 2007. Appendix 2, on the other hand, compares change between PISA 2003 and 2009 against change between TIMSS 2003 and 2007.

[13] The mean and standard deviation used in the z-score transformation are based on data weighted by country size. Brown et al (2007), on the other hand, use an unweighted average.

[14] Each of the ten countries took part in the PISA 2000, PISA 2009, TIMSS 1999 (8th grade) and TIMSS 2007 (8th grade) survey waves.

Figure 2

There seems to be broad agreement between the two surveys. Both identify Japan, Hong Kong and Korea as high maths test performers while Indonesia is at the other end of the scale. The six other countries (including England) are bunched somewhere in-between, with exact positions within this sub-group slightly less clear. The correlation between estimates is nevertheless high (r = 0.93 including Indonesia and 0.83 without), with England sitting almost exactly on the 45 degree line. In analysis not presented here (for brevity) similar results held for selected points of the test distribution (e.g. the 25th and 75th percentile). It hence seems that the latest PISA and TIMSS survey waves provide a reasonably consistent picture of where England currently stands within this group of countries (Appendix 1 shows that this remains true if a number of other developed nations are also included in the analysis).

What the two studies disagree on, however, is how the average performance of English pupils has changed over time. This is clearly illustrated in Figure 3, where the average test score for England (in terms of international z-scores) is plotted for each survey wave since 1999 (TIMSS – solid black line) or 2000 (PISA – dashed black line). Within this fixed pool of ten countries, PISA test scores have declined over this period (from a z-score of over 0.40 in 2000 to one of around 0.20 in 2009). Yet, in the TIMSS data, the exact opposite holds true (the average z-score has increased from below 0 in 1999 to just over 0.20 in 2007)[15].

Figure 3

Further detail is provided in Table 1 where the distribution of test scores is presented for England from the 1999 and 2007 TIMSS and 2000 and 2009 PISA survey waves. This reveals whether the inconsistency between PISA and TIMSS is specific to one part of the test distribution (e.g. whether the inconsistency lies more in the bottom than in the top, or vice-versa). Recall that all figures refer to international z-scores.

[15] This conflict between the two surveys has been commented on briefly by others (Smithers 2011), with the change in PISA and TIMSS ranking being the subject of a major debate between Labour and Conservative MPs. See, for instance,
http://www.publications.parliament.uk/pa/cm201011/cmpublic/education/110324/am/110324s01.htm

Table 1

At each of the 10th, 25th, 75th and 90th percentiles the two surveys tell a conflicting story about England's maths performance over time – PISA suggests it is going down and TIMSS that is going up. It is, however, interesting to also consider the measures of spread in the bottom half of the table. The two surveys seem to agree that there has been little overall change in educational inequality between 2000 and 2009 as measured by either the standard deviation or difference between the 90th and 10th percentile. Looking at the P90-P50 comparison, however, PISA suggests there has been some increase within the top half of the test distribution, while in TIMSS there is evidence of a decline. Both studies, on the other hand, agree that the gap between the 10th and 50th percentile has increased – although there is some conflict in the extent to which this has occurred. Consequently, there is some suggestion that PISA and TIMSS also disagree about how inequality in educational achievement may have changed over this period.

It is hence clear that these two major international studies conflict on how secondary school children's maths test scores have changed over time. What is perhaps even more intriguing, however, is that this inconsistency is not an experience shared by other countries. Evidence is presented in Figure 4, where the change in mean PISA maths test scores between 2000 and 2009 is plotted on the x-axis, with the change for mean TIMSS scores between 1999 and 2007 on the y-axis. The 45 degree line illustrates where results from the two studies "agree" (i.e. where the estimated change in PISA is equal to that in TIMSS).

Figure 4

One can see that most countries are scattered reasonably tightly around this line, with the change observed in TIMSS similar to that in PISA, typically differing by 0.10 of an international standard deviation or less[16]. England, however, is an obvious outlier. The difference between the change observed in the PISA and TIMSS surveys is around half an international standard deviation – double that in any other country and five times greater

[16] The difference for Italy is slightly bigger at around 0.20 of an international standard deviation. The PISA and TIMSS surveys do, however, agree on the direction of change for Italy, if not the exact magnitude (unlike the results for England).

than what is "typical". This could, of course, just be a matter of sampling variation. To investigate this possibility, a two-sample t-test (assuming independent samples) has been conducted (see Appendix 3 for further details). The null hypothesis (that the change in mean test scores is the same across the two studies) cannot be rejected in eight of the ten countries considered. In Italy and England I can reject this null at the 1% level, although the t-statistic is almost half the size in the former (t = 3.0) than it is in the latter (t = 5.8). Sampling variation could, of course, still explain some of the difference between the two surveys. Yet it is also clear that other factors (e.g. possible non-sampling error) may be at play.

4. The comparability of the international achievement data for England over time

This section discusses three issues with regard to the comparability of the international achievement data for England across the survey waves - the target population, survey procedures and non-response. This list may not be exhaustive, but rather draws upon my research into the data and experience in their use.

4.1 Target population

There seems to have been at least two changes to the target population between the PISA 2000 and 2009 survey waves (I am not aware of any similar changes in TIMSS). The first is that in the 2000 wave the sample included just children from England and Northern Ireland. But, from 2003 onwards, it also included young people from Wales. This could be problematic as figures from the latest PISA study illustrate that Welsh pupils have significantly lower levels of achievement (Bradshaw et al 2010 show that Welsh pupils scored an average of 472 on the PISA 2009 maths test compared to 492 for those from England). The fact that Welsh schools did not take part in the PISA 2000 study hence means that the average PISA maths test score for "England" in that year is likely to be higher than in the other survey waves[17]. This, in turn, means that there is also potential overestimation of change over time. How much impact does this have on the substantive finding that PISA test scores for England have declined? This is shown via the dotted line in Figure 3, where the PISA trend since 2000 has been re-estimated having excluded Welsh schools from the

[17] This is essentially because a group of low achieving children have not taken part in the 2000 study.

2003, 2006 and 2009 analysis (and thus restricting focus to England only). Clearly the impact of this is minimal, with the pronounced decline in test scores remaining[18].

The second change to the target population is that the PISA data for England have been altered from an age-based sample in 2000 and 2003 to what is (for all intents and purposes) a grade-based sample in 2006 and 2009. In other words, students in the older PISA cohorts were all born in the same calendar year (1984 in the case of PISA 2000 and 1987 in the case of PISA 2003), with roughly two-thirds of children in "year 11" and one third in "year 10". On the other hand, almost all the children who sat the PISA test in 2006/2009 all belonged to the same academic year (i.e. almost all the PISA 2009 participants were year 11 students born between September 1993 and August 1994). Moreover, my exploration of the data suggests that this is something that did not occur in other countries (i.e. it is a specific change made to the PISA study in England)[19]. Appendix 4 and Appendix Table 1 provide further details.

What impact does this have on my results? To provide some indicative evidence on this issue, mean test scores for England are re-calculated having restricted the sample to year 11 pupils who are born between January and August in all four survey waves. This leads to a slight increase in scores for the two earliest rounds of the survey (the mean international z-score for England increases from 0.43 to 0.47 in 2000 and from 0.35 to 0.39 in 2003) and a slight decrease in the later rounds (the mean z-score for England drops from 0.27 to 0.26 in 2006 and from 0.23 to 0.22 in 2009). In other words, the decline in England's PISA test scores over time may have been underestimated because of this issue. However, caution is required when interpreting this result as other changes have been made to the conduct of the PISA study over the same period. These are detailed in the following sub-section.

[18] The reason is that the PISA survey weight scales observations up to the national population. Welsh children (making up only a relatively minor proportion of the "England and Wales" population) receive little weight in the calculation of average test scores in the 2003, 2006 and 2009 PISA studies. So when they are dropped from the analysis, there is little change in the results (at least in terms of average test scores).

[19] This change between using an age and grade based sample is not something that has been explicitly documented (to my knowledge) in either the national or international report. Rather this is based upon my investigation of the month of birth, year of birth and grade attended variables that are part of the international database (downloaded from the PISA website www.pisa.oecd.org). Appendix 4 provides further details.

4.2 Survey procedures

Whereas the first two PISA waves for England were conducted by the UK national statistics agency (the Office for National Statistics), the 2006 and 2009 studies were contracted out to an external provider (the National Foundation for Educational Research). This seems to have been accompanied by some changes to the survey procedures (Appendix 4 provides specific details)[20]. Perhaps the most important (that I am aware of) is that the month when children sat the PISA test moved from between March and May (in PISA 2000 / 2003) to roughly five months earlier (November/December) in PISA 2006 / 2009. England had special dispensation to make this change (i.e. this is not something that occurred in other countries) and although this was for good reason (the PISA 2000 and 2003 studies clashed with preparation for national exams and so was a significant burden on schools) it may have had unintended consequences. Again, I believe this is a change that has occurred only in PISA and not in TIMSS.

How might this influence the trend in England's PISA test scores? Firstly, it is important to understand that between November/December and March-May of year 11 is likely to be a period when children add substantially to their knowledge of the PISA subjects as it is when pupils are working towards important national exams. Consequently, one should expect the year 11 pupils in the PISA 2000/2003 cohort to out-perform their peers taking the test in 2006/2009 due to the extra five months they have had at school. To provide an analogy, imagine that one group of children took a mock GCSE maths exam in November, and another group the following April; clearly one would expect the former to obtain lower marks (on average) than the latter. This would in turn suggest an overestimation of the decline in PISA maths scores over time[21]. Putting a figure on the size of this potential bias is not easy, although it has been widely cited that one additional school year is equivalent to roughly 40 PISA test points (0.4 of an international standard deviation).

[20] It should be noted that the change of survey procedures had nothing to with the NFER taking over the PISA contract per se. Rather it was part of the strategy to raise England's disappointing response rate in the 2000 and 2003 waves.

[21] One could, of course, put forward the counter-argument that children towards the end of year 11 are so focused on exam performance that they add little to their "functional knowledge" which is tested in PISA. Micklewright et al (2010) and Micklewright and Schnepf (2006) show, however, that children's PISA test scores correlate very highly with national exam performance (the correlation between key stage 3 and PISA maths scores is over 0.8) suggesting that success in one is closely related to that on the other.

See OECD (2010b page 110) for further details[22]. This would imply that year 11 children who sat the PISA test in 2000 might be expected to outperform the 2009 cohort by roughly 15 PISA test points (0.15 of an international standard deviation) due to their additional five months at school[23].

4.3 Non-response

It has been widely recognised that non-response is a problem for England in the international achievement datasets (and more so than in most other countries), although discussion of this issue has mainly focused upon PISA (OECD 2010a, Micklewright et al 2010, Micklewright and Schnepf 2006)[24]. In fact, this was the reason given by the OECD for excluding England from a recent report on changes in PISA test scores over time. Specifically, they state that[25]:

"The PISA 2000 and 2003 samples for the United Kingdom did not reach response-rate standards, so data from the United Kingdom are not comparable to other countries."
(OECD 2010 page 26)

Interestingly, however, they add a footnote saying that (with regard to the 2000 data):

"the PISA consortium concluded that response bias was likely negligible"

(OECD 2010a page 30 note 3).

If this is indeed true, then it seems unlikely that missing data in PISA would substantially bias any comparison of England's performance in 2000 with that in 2009.

[22] This has become a widely cited statistic by users of the PISA data, including both academics and policymakers. It stems from page 110 of OECD (2010b) See http://browse.oecdbookshop.org/oecd/pdfs/free/9810091e.pdf for further details. The OECD has since used this statistic in their own publications (see 2010c – page 157). An unpublished note provided by England's Department for Education supports this assumption.

[23] If 40 PISA points is equivalent to an additional year of schooling, then one month is equivalent to around 3 points. The five month difference in schooling between the PISA 2000 and 2009 cohort is thus the equivalent of around 15 PISA test points.

[24] This is probably due to the fact that the UK was excluded from the PISA 2003 international report because of missing data problems in England.

[25] The international report groups England, Scotland, Wales and Northern Ireland together as the UK. However, although the OECD discusses response rates with regard to the UK as a whole, it was actually the sample for England that did not meet the criteria set.

Yet other studies suggest that this may not be such a trivial issue. Micklewright et al (2010), for instance, used English children's administrative records (including information on their national exam scores) to investigate non-response bias in the 2000 and 2003 PISA data[26]. They concluded that the average maths test score for England in the 2000 wave was upwardly biased by somewhere between 4 and 15 points (page 33 Table 7b), with their preferred estimates towards the top of this scale. Assuming that this problem was confined to the PISA 2000 and 2003 studies (i.e. non-response had a negligible impact on the average test score for England in 2006 and 2009) then this by itself could explain a large part of the decline seen in England's PISA test scores over the past decade.

This is, however, of only limited use to address the issue at hand. To better understand change over time, one ideally needs to know (a) how the bias for England has changed between each of the four PISA survey waves and (b) if there is similar bias in the TIMSS data. Unfortunately, there has been little work addressing these issues (and is, to my knowledge, not currently being planned)[27]. It is possible, however, to investigate how the response rate has changed over time. Improving (or higher) response rates does not, of course, mean that there will necessarily be less bias, but nevertheless provides some guidance on this issue. Details are provided in Table 2 below.

Table 2

There is some evidence of improving response rates in PISA over time. This has, however, been reasonably modest, with the percentage of schools taking part (before replacement schools are considered) increasing by ten percentage points between 2000 and 2009 (from 59% to 69%) with pupil response going up by around six percentage points (from 81% to 87%). If this has reduced the upward bias in mean test scores found in the earlier PISA waves (by Micklewright et al) then this may explain some of the decline in England's performance over this period. But as one is unable to also investigate the pattern of

[26] The authors create a set of response weights for the English data based upon this rich auxiliary information. This potentially allows the authors to make a better correction for non-response bias than is possible with the weights supplied by the OECD in the international database.

[27] Bradshaw et al (2010a) does state that, because the UK data fell short of the OECD target participation rate in PISA 2009, it had to provide some analysis of the characteristics of responding and non-responding schools. They report no significant differences, and so suggest there was no evidence of possible bias. However these details were only provided to the PISA sampling referee and have not been made publicly available. Without such information, it is difficult to comment further on this issue.

response in 2006 and 2009, there remains some ambiguity over the extent to which this can explain the trends presented in Figures 2 and 3.

The problem of missing information has received rather less attention in TIMSS. Panel b of Table 2 suggests, however, that less than half of the first choice schools in 1999 (49%) and 2003 (40%) agreed to take part. This changed, however, in 2007 when participation reached near 80%. This has important implications for the interpretation of the TIMSS trend for England in Figure 3 – the doubling of the school response rate coincides with the marked improvement in average test scores (i.e. the big increase from 2003 to 2007). However, without more information on the nature of this non-response (and how it has changed over time) it is again difficult to decipher whether England's rise up the international rankings in TIMSS is an artefact of the data or represents genuine change.

4.4 The cumulative impact on the trend in average PISA maths test scores

A number of difficulties have been identified with the PISA data for England. But what is the cumulative impact of these on the PISA trend shown in Figure 3? Four estimates, based upon different assumptions about the comparability of the data across survey waves, are now produced. These can be summarised as follows:

- Estimate 1 – No adjustment is made to the raw PISA test scores. In other words, one ignores the issues discussed in this section and assumes that data from the four waves are already comparable (solid black line).
- Estimate 2 – Only English year 11 pupils born between January and August are included in the analysis (i.e. Welsh and year 10 pupils are dropped) so that the target population is consistent between the different waves. No adjustment is made, however, for the change of survey month or difficulties with non-response (dashed grey line with circle markers).
- Estimate 3 – As estimate 2, but with test scores lowered by 15 points for the 2000 and 2003 sample members to account for the fact that these children would have

had five months more tuition when they took the PISA test (dashed grey line triangular marker)[28].

- Estimate 4 - As estimate 3, but with mean test scores lowered by a further 15 points in 2000 and 7 points in 2003 to account for the non-response bias found in the Micklewright et al study (grey dotted line, cross as markers)[29,30.]

Results can be found in Figure 5.

Figure 5

The trend varies substantially depending upon the assumptions made. For instance, there is a decline of 0.25 of an international standard deviation in "estimate 2", but a small rise of 0.05 in "estimate 4". This clearly brings into question whether the performance of secondary school pupils performance in England has really been in relative decline. It would be wrong, however, to claim that any one of the four estimates is "correct", or that the TIMSS data should be used instead[31]. Rather the key point is that there are problems with identifying change over time using the PISA data for England, and that conclusions (and public policy) must not be based on this resource alone. Indeed, given that other evidence (from TIMSS and national exam results) contradicts PISA, it is difficult to treat the apparent decline in secondary school pupils' performance as a statistically robust result.

5. Conclusions

The international studies of pupil achievement provide an important insight into how secondary school children's achievement varies across countries and is changing over time. Policymakers in England have paid much attention to the PISA data with regard to this issue, but are results from this single study "robust"? This paper has shown how the PISA and TIMSS data for England are problematic, and that they do not provide a clear and consistent

[28] The figure of 15 PISA test points is explained in section 4.2. This is incorporated into the analysis by subtracting 15 points from the 2000 and 2003 sample members' raw PISA test scores and then re-running all estimations.

[29] These figures are based on the GREG weights produced by Micklewright et al (page 32/33 Table 7a and 7b) as discussed in section 4.3. These weights would ideally be used in this analysis but are not publicly available. Instead, the mean PISA maths test score for England in 2000 and 2003 is lowered by the relevant amount.

[30] An implicit assumption here is that the estimates of average PISA test scores for England in the 2006 and 2009 samples are unbiased. This may, of course, not be the case. But as explained in section 4.3, without further data to investigate this issue, this is the most logical assumption to make.

[31] Indeed, Figure 5 clearly shows that I am unable to rule out that secondary school pupils' performance may have declined over the past decade as suggested by the raw PISA results.

message about how secondary school children's performance has changed in this country (relative to others). There are specific problems with missing data, survey procedures and the target population, which limit the inferences one can draw. The recommendations made to policymakers are therefore as follows:

- One cannot firmly conclude that English secondary school children's performance has improved or declined relative to that of its international competitors over the past decade.
- The decline seen by England in the PISA international rankings is not, in my opinion, statistically robust enough to base public policy upon.
- The decline in PISA test scores does not suggest that the Labour government's investment in education was a waste of money, just as the ascendency in the TIMSS rankings does not prove it was well spent[32].

There are also some clear practical messages for England's policymakers and international survey organisers to take from this paper. The first is that better documentation of the issues discussed is needed, both in the national and international reports. Secondly, it may be possible to get a better understanding of England's comparative performance over time by linking the international achievement datasets to children's administrative records. England is somewhat unusual in having national measures of performance collected at ages 7, 11, 14 and 16 and this could potentially be exploited to investigate at least some of the issues raised. Indeed, if policymakers want to continue to make statements on this issue, then such data linkage should be strongly considered. Thirdly, researchers using such data to investigate trends over time in England should make readers aware of the issues discussed in this paper and check the robustness of their results. This might include an investigation of whether consistent results are obtained from different datasets (e.g. that their results hold in both PISA and TIMSS) or with other research. Finally, although response rates in England for PISA and TIMSS have improved, there is still a struggle to meet international standards. Not enough information is provided to users on how this may influence their results. In future waves, data linkage and bias analysis (with results fully documented in the national report) should be undertaken as a matter of course. Moreover,

[32] Indeed, even if the data were of high enough quality to accurately estimate changes over time, such statements seem to fall into the trap of confusing correlation with causation.

the production of additional material to help correct for any of the problems discovered (e.g. response weights) should be made publicly available.

References

Bradshaw, J. Ager, R. Burge, B. and Wheater, R. (2010) "PISA 2009: Achievement of 15-year-olds in England", Slough: NFER

Bradshaw, J. Ager, R. Burge, B. and Wheater, R. (2010) "PISA 2009: Achievement of 15-year-olds in Wales", Slough: NFER

Bradshaw, J. Sturman, L. Vappula, H. Ager, R. and Wheater, R. (2007a) "Achievement of 15-year-olds in England: PISA 2006 National Report" , Slough: NFER

Bradshaw, J. Sturman, L. Vappula, H. Ager, R. and Wheater, R. (2007b) "Achievement of 15-year-olds in Wales: PISA 2006 National Report", Slough: NFER

Brown, G. Micklewright, J. Schnepf, S. and Waldmann, R. (2007) "International Surveys of Educational Achievement: How Robust Are the Findings?" Journal of the Royal Statistical Society Series A, volume 170, issue 3, pp. 623-46

Crawford, R. Emmerson, C. and Tetlow, G. (2009) "A Survey of Public Spending in the UK", IFS briefing note BN43

Department for Children, Schools and Families (2009) "Departmental Report 2009", London Downloadable from http://publications.dcsf.gov.uk/eOrderingDownload/DCSF-Annual%20Report%202009-BKMK.PDF

Gill, B. Dunn, M. and Goddard, E. (2002), "Student Achievement in England: Results in Reading, Mathematical and Scientific Literacy Among 15-year-olds from OECD PISA 2000 study", Office for National Statistics Report

Goldstein, H. (2004b) "International Comparisons of Student Attainment: Some Issues Arising From the PISA study", Assessment in Education, volume 11, pp 319-330

Martin, M. Gregory, K. & Stemler (2000) "TIMSS 1999 Technical report", IEA

Micklewright, J. and Schnepf, S. (2006) "Response Bias in England in PISA 2000 and 2003", Department for Education and Skills Research Report 771

Micklewright, J. Schnepf, S. and Skinner, C. (2010) "Non-response Biases in Surveys of School Children: The Case of the English PISA Samples", DoQSS Working Papers 10-04

OECD (2009) "PISA DATA ANALYSIS MANUAL: SPSS SECOND EDITION", Paris: OECD

OECD (2010a) "Learning Trends: Changes in Student Performance since 2000", Paris: OECD

OECD (2010b) "Learning to Learn: Student Engagement Strategies and Practices Volume III", Paris: OECD

OECD (2010c) "PISA 2009 results: What Students Know and Can Do", Paris: OECD

Ruddock, G. Sturman, L. Schagen, I. Styles, B. Gnaldi, M. and Vaoula, H. (2004) “Where England Stands in the Trends in International Mathematics and Science Study (TIMSS) 2003”, NFER

Smithers, A. (2011) “The Inconvenient Truth About the Global School League Tables”, Parliamentary brief online

Smithers (2011b), “GCSE 2011”, Centre for Education and Employment Research report

Sturgis, P, Smith P and Hughes G (2006) “A Study of Suitable Methods for Raising Response Rates in School Surveys”, Department for Education and Skills Research Report 721

Sturman, L. Ruddock, G. Burge, B. Styles, B. Lin, Y. and Vappula, H. (2008) “England’s Achievement in TIMSS 2007: National Report for England”, NFER

Wu, M. (2010) “Comparing the Similarities and Differences of PISA 2003 and TIMSS”, OECD Education Working Papers No. 32

Table 1. Distribution of test scores for England in the PISA 2000 and 2009 and TIMSS 1999 and 2007 survey waves (international z-scores)

	PISA			TIMSS			Difference in change between the two surveys
	2000	2009	Change	1999	2007	Change	
P10	-0.61	-0.86	-0.25	-1.08	-0.85	0.23	0.48
P25	-0.09	-0.34	-0.24	-0.61	-0.30	0.30	0.55
P50	0.47	0.24	-0.23	-0.08	0.28	0.35	0.58
Mean	0.43	0.23	-0.20	-0.07	0.24	0.31	0.51
P75	0.99	0.78	-0.21	0.46	0.82	0.36	0.57
P90	1.48	1.34	-0.13	0.95	1.25	0.29	0.42
SD	0.82	0.84	0.02	0.80	0.80	0.00	-0.02
P90 - P10	2.09	2.21	0.12	2.03	2.09	0.06	-0.06
P90 - P50	1.01	1.10	0.09	1.03	0.97	-0.06	-0.16
P50 - P10	1.08	1.10	0.03	1.00	1.12	0.12	0.10

Notes:

Figures are reported in terms of international z-scores

Table 2. School and pupil response rates in the PISA and TIMSS datasets

(a) PISA

Year	Source	School		Pupil
		Before replacement	After replacement	
2000	Micklewright & Schnepf (2006)	59	82	81
2003	Micklewright & Schnepf (2006)	64	77	77
2006	Bradshaw et al (2007a)	77	89	89
2009	Bradshaw et al (2010a)	69	87	87

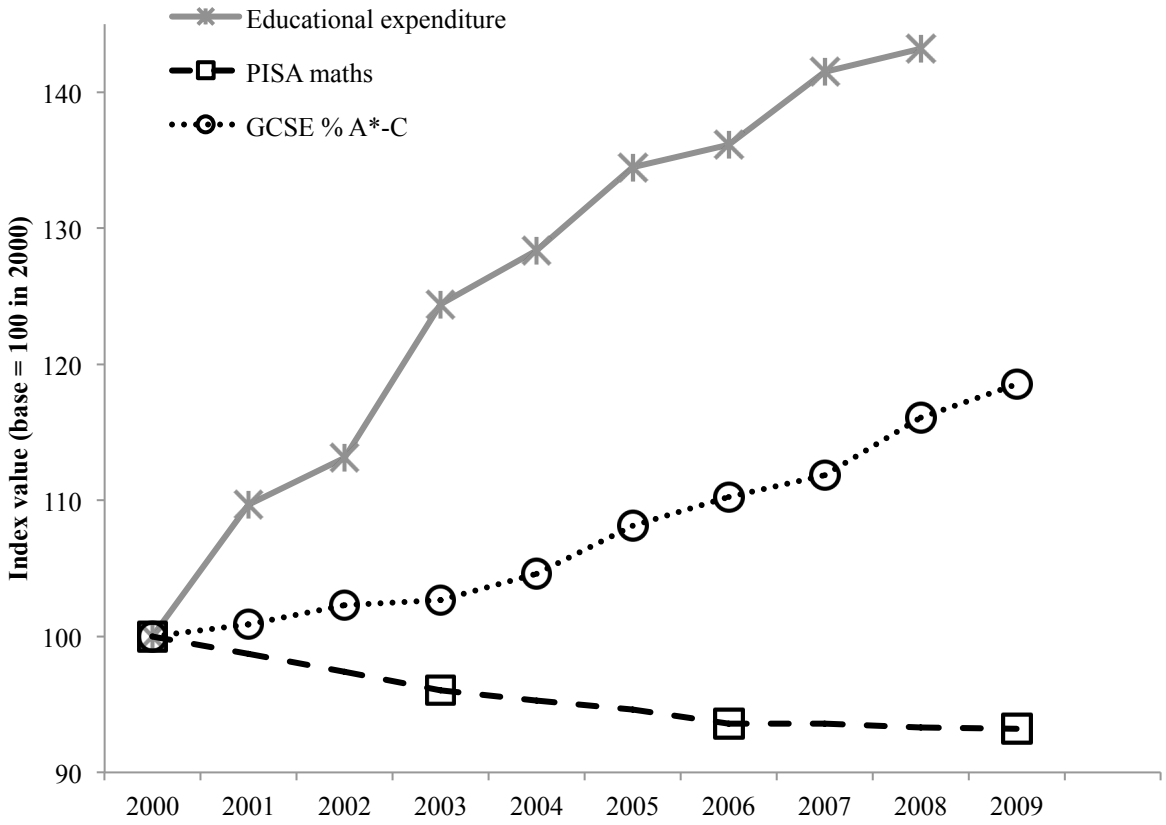
(b) TIMSS 8th grade

	Source	School		Pupil
		Before replacement	After replacement	
1999	Martin et al (2000)	49	85	90
2003	Ruddock et al (2004)	40	54	86
2007	Sturman et al (2008)	78	86	88

Notes:

Figures refer to percentage of schools / children who agree to take part in the study. After replacement refers to total percentage of schools who agree to take part after first and second replacements have been included.

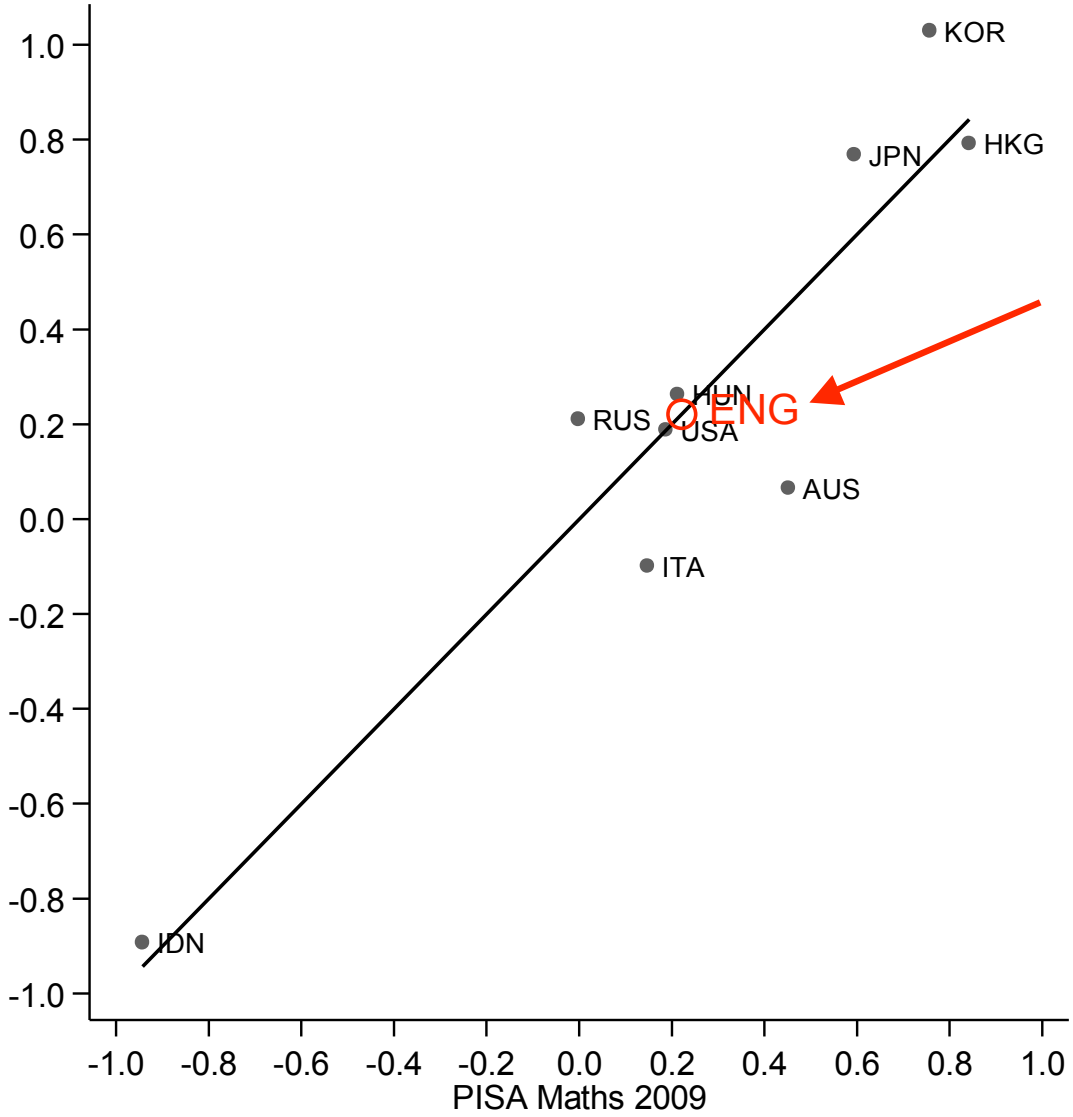
Figure 1. Change in real educational expenditure and mean PISA maths test scores in England between 2000 and 2009



Notes:

Figures have been based to an index value of 100 in the year 2000. The solid line refers to the trend in educational expenditure since 2000, the dashed line represents the trend in children's PISA scores and the dotted line the proportion of children achieving at least 5 A*-C in their national (GCSE) exams. Data on educational expenditure drawn from Department for Children, Schools and Families (2009) Table 8.5, page 177, third row down (labelled "current"). These figures refer to *current* expenditure on under 5, primary and secondary education and excludes administration costs. PISA test scores are the author's calculations based upon the PISA international database. GCSE scores are taken from Smithers (2011b) page 2, chart 1.2 (http://www.buckingham.ac.uk/wp-content/uploads/2010/11/GCSE2011.pdf)

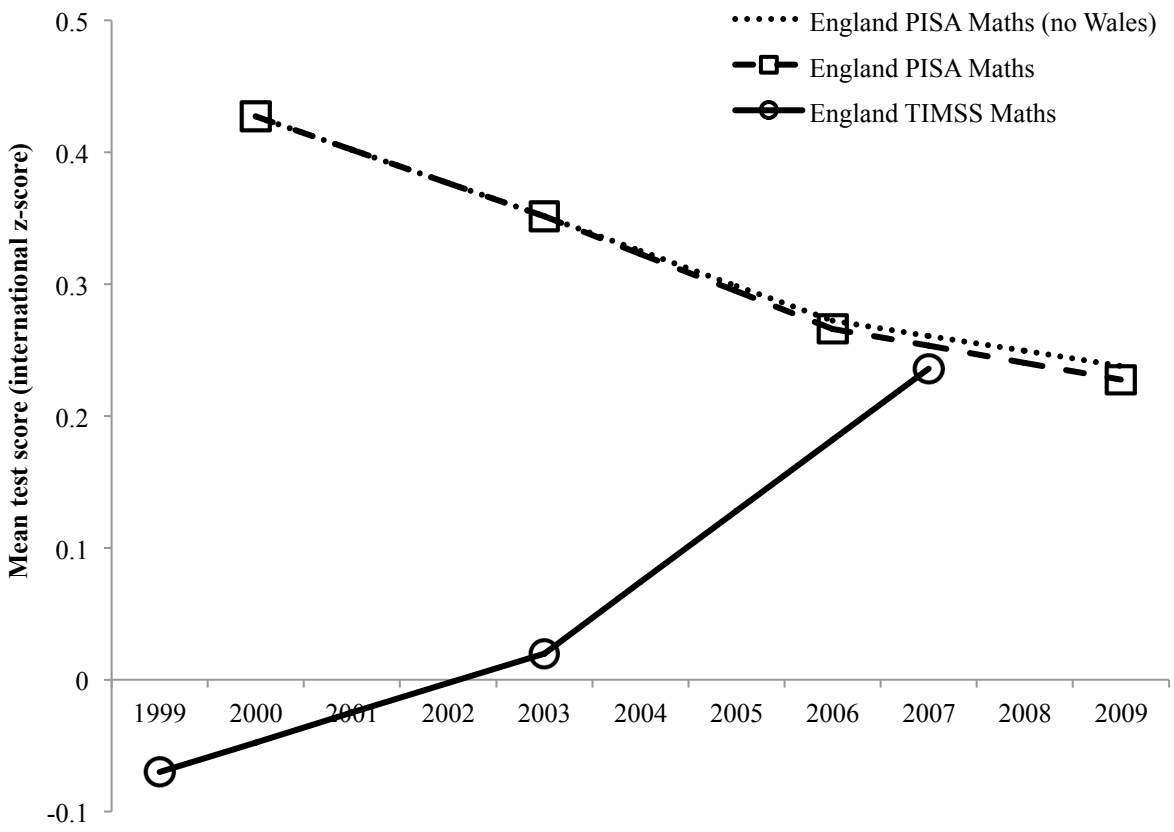
Figure 2. Results for TIMSS 2007 and PISA 2009: mean maths scores for a pool of 10 countries



Notes:

Figures are presented in terms of international z-scores, with the data having been standardised within the sub-set of the 10 countries considered. PISA 2009 maths test scores sit on the x-axis while TIMSS 2007 scores run along the y-axis. The solid 45 degree line represents where mean test scores in PISA are the same as those for TIMSS.

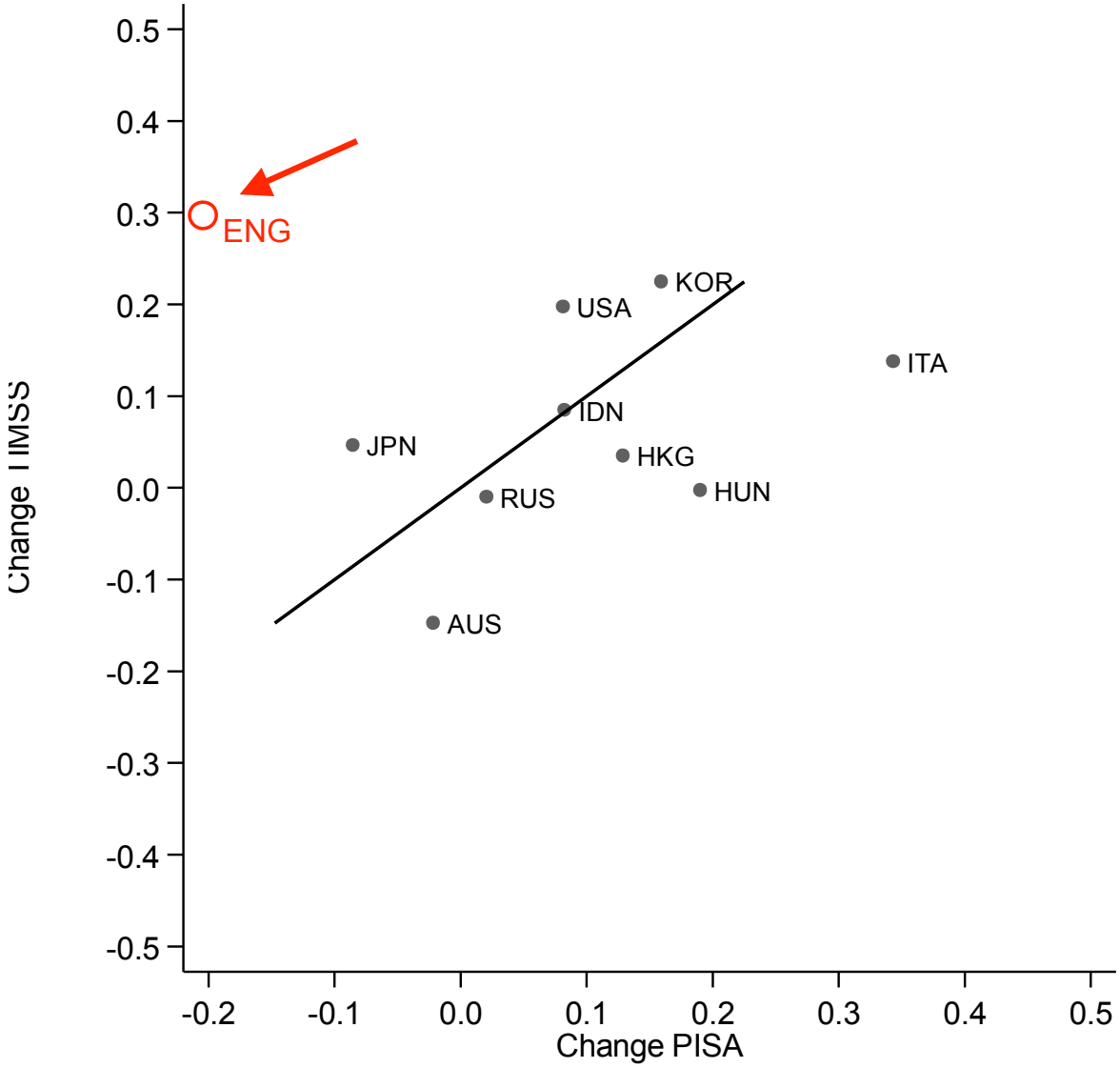
Figure 3. Change in PISA and TIMSS (8th grade) maths test scores over time



Notes:

The black dash line refers to PISA maths test scores for England between 2000 and 2009. The dotted line refers to when one excludes children in Welsh schools from PISA. The solid line, on the other hand, refers to TIMSS maths scores between 1999 and 2007. Figures presented on the y-axis refer to the average test performance and are presented in terms of international z-scores.

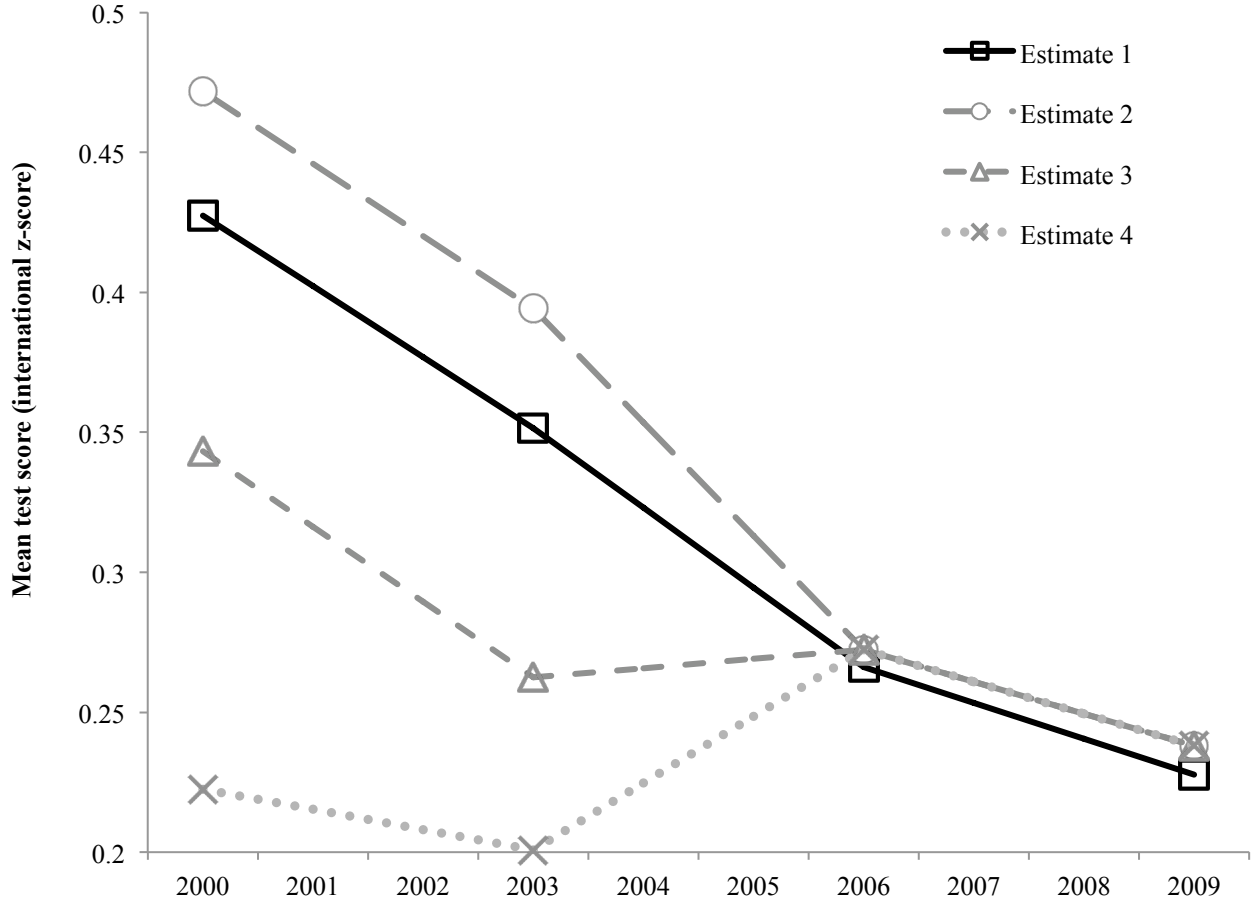
Figure 4. Change in PISA mean maths scores 2000-2009 compared to mean TIMSS maths scores 1999 – 2007



Notes:

Figures on the x-axis refer to the change in mean PISA test scores between 2000 and 2009. Those on the y-axis, on the other hand, refer to the change in TIMSS scores between 1999 and 2007. All figures presented are in terms of international z-scores. The solid black line represents where the change in PISA test scores over the period is the same as the change in TIMSS test scores.

Figure 5. Alternative estimates of the trend in mean PISA maths test scores for England



Notes:

These estimates are based on four different sets of assumptions that are discussed in section 4.4. Estimate 1 is the trend for England based on the raw PISA data. Estimate 2 is where the data is restricted to just year 11 pupils born between January and August in England. Estimate 3 is the same as estimate 2, but when an adjustment has been made for the change of test month. Estimate 4 is the same as estimate 3, but when an additional adjustment has been made for non-response bias in the 2000 and 2003 PISA waves.

Appendix 1. A comparison of results from PISA and TIMSS using a wider array of countries

In the main body of this paper the data are restricted to ten countries that took part in each of the 2000, 2003, 2006, 2009 PISA and 1999, 2003 and 2007 TIMSS waves. A greater number of countries is now used, based on more lenient criteria regarding which rounds the countries have participated in. Specifically, the Czech Republic, the Netherlands, New Zealand, Norway, Scotland and Sweden are now included in the analysis (in addition to the ten countries in section 3). As Norway, Sweden and Scotland did not take part in TIMSS 1999, when estimating change over time TIMSS 1995 data is used in its place[33]. Similarly, New Zealand and the Netherlands do not have TIMSS 2007 data, so for these countries estimates of change over time refer to the difference between mean maths test scores in 1995 and 2003.

Before considering the issue of change, the cross-sectional picture is presented from TIMSS 2007 and PISA 2009. In other words, do these studies agree on how England's mean test performance in the two most recent survey waves compares? Estimates are presented in terms of international z-scores and can be found in Appendix Figure 1 (these are analogous to the estimates in Figure 2 – but now using a wider selection of countries). The y-axis refers to results from TIMSS 2007, while those on the x-axis come from PISA 2009.

Appendix Figure 1

As discussed in section 3, England is behind a set of leading countries in both surveys (Hong Kong, Japan, Korea, Netherlands), while being clearly ahead of one (Indonesia) with a particularly low level of income. More generally, there seems to be reasonable agreement between the PISA and TIMSS results (the correlation coefficient equals 0.85 including Indonesia and 0.74 without). Yet there is also a cluster of developed countries between the two extremes where there is some conflict. Australia and New Zealand, for instance, both seem have higher mean maths test performance on the PISA 2009 test (average international z-score ≈ 0.4) than TIMSS 2007 (average international z-score ≈ 0). England, however, sits almost exactly on the 45 degree line. This suggests that the most recent waves

[33] In other words, for these countries change over time in TIMSS is calculated as the difference in mean scores between 1995 and 2007 rather than between 1999 and 2007.

of the PISA and TIMSS studies agree about how England's secondary school children are performing in maths relative to their peers in other countries.

Appendix Figure 2 turns to the issue of change over time. The change in mean PISA maths test scores on the x-axis (2000 - 2009) with the change in TIMSS (1999 – 2007) on the y-axis can be found[34]. The 45 degree line illustrates where results from the two different studies "agree" (i.e. where the estimated change in PISA is equal to that in TIMSS).

Appendix Figure 2

As discussed throughout this paper, the TIMSS and PISA data strongly disagree about how England's secondary school pupils maths test performance has changed since the turn of the 21st century. The former suggests that there has been a notable improvement, while in the latter there has been a notable decline. Moreover, as previously suggested in Figure 4, England is unusual in this respect compared to other countries (one can see it is the biggest outlier in Appendix Figure 2). This result still holds, even now a wider selection of countries has been included[35]. Together, this further supports the argument that England is unusual in the extent to which PISA and TIMSS disagree, and that this may be due to data difficulties (in either or both surveys) for this particular country.

Appendix 2. Do results change when one uses different subjects or different survey years?

In the main body of this paper I have chosen to compare each country's change in maths test performance over time. As noted in section 2, however, there are some difficulties with this approach (e.g. the maths scale in PISA was not fully developed until the 2003 wave). This Appendix investigates whether findings are robust to the use of different subject areas instead. In other words, is the disagreement between PISA and TIMSS in England's test performance over time specific to maths, or does it hold no matter what subject is used?

[34] Due to data limitations (as described above) the TIMSS data for New Zealand and Netherlands is based upon the 1995 and 2003 waves, while for Norway, Scotland and Sweden it refers to the 1995 to 2007 comparison.
[35] The Netherlands also stands out in Appendix Figure 2. However, the response rate of first choice schools was very low (23%) in the base year (1995) but improved markedly by 2003 (79%). This improvement in school response over time is similar to that seen in England.

There are six possible PISA-TIMSS subject combinations that can be investigated. These are:

- The change in PISA maths scores VS The change in TIMSS maths scores
- The change in PISA reading scores VS The change in TIMSS maths scores
- The change in PISA science scores VS The change in TIMSS maths scores
- The change in PISA maths scores VS The change in TIMSS science scores
- The change in PISA reading scores VS The change in TIMSS science scores
- The change in PISA science scores VS The change in TIMSS science scores

Estimates for each of the above are presented in Appendix Figure 3. This is in the form of a 3x2 matrix, with each graphic providing an analogous set of estimates to those in Figure 4, but using a different combination of PISA-TIMSS subjects. Labels running horizontally and vertically along the edge of the page identify which have been used. England is highlighted in red with a hollow circular marker throughout.

Appendix Figure 3

England is the furthest country from the 45 degree line in each of the diagrams, with one able to reject the null hypothesis that the change in PISA is the same as the change in TIMSS on every occasion. There is some suggestion, however, that England is slightly less of an outlier when looking at change in science scores rather than maths (i.e. the figure in the bottom right). The general finding that there is a discrepancy between the change observed in PISA and TIMSS holds no matter which subject is used.

Attention turns to the choice of survey years in Appendix Figure 4.

Appendix Figure 4

As noted in section 2, the PISA maths score was technically not fully developed until the 2003 wave. Do results still hold if the 2003 wave is used as the base year rather than 2000? Appendix Figure 4 suggests that this is indeed the case. England is still an obvious outlier, with a notable difference in the change between 2003 and 2009 in PISA than the change between 2003 and 2007 in TIMSS.

Appendix 3. Comparison of change over time in PISA to the change over time in TIMSS: calculation of the estimated standard errors

I begin by calculating the difference in mean test scores for each country between the two waves. This provides an estimate of change over time. I do this twice for each country, once using the PISA data (comparing the 2000 and 2009 wave) and once using TIMSS (comparing the 1999 to 2007 waves). Samples are assumed to be independent between survey waves, with standard errors for the change over time (within each survey) calculated using the formula:

$SE_{CHANGE} = \sqrt{(SE_{T0}^2 + SE_{T1}^2)}$

Where

SE_{CHANGE} = Standard error of the change in mean test scores between the two survey waves

SE_{T0} = Standard error of mean test scores in the earlier survey wave (i.e. either PISA 2000 or TIMSS 1999)

SE_{T1} = Standard error of mean test scores in the later survey wave (i.e. either PISA 2009 or TIMSS 2007)

I then compare the PISA and TIMSS estimates of change over time. These are the figures presented along the x-axis and y-axis of Figure 4 and Appendix Figure 2. To investigate whether the change seen in PISA is significantly different from that in TIMSS, I conduct the following two sample t-test (assuming independent samples) for each country:

$$T_{STAT} = \frac{(CHANGE\ PISA - CHANGE\ TIMSS)}{\sqrt{SE\ CHANGE_P^2 - SE\ CHANGE_T^2}}$$

Where:

$SE\ CHANGE_P$ = Standard error of the change in mean test scores between the 2000 and 2009 PISA waves

$SE\ CHANGE_T$ = Standard error of the change in mean test scores between the 1999 and 2007 TIMSS waves

Appendix 4. Details on the change of test month and target population

In this appendix I provide further information on the PISA sample for England. Four topics are covered:

- Welsh participation in PISA
- Identification of pupils from England in the international database
- References for the change of test month
- Further information on the move from an age to grade based sample

Welsh participation in PISA

Gill (2002 Appendix B page 165) states that:

"Wales did not participate in PISA 2000."

Yet in future rounds of the study (e.g. in 2006 and 2009) a national report for Wales has been published (e.g. Bradshaw et al 2007b).

Identification of pupils from England in the international database

The "country" and "subnatio" variables can be used in PISA 2000 to identify children from England. If the data are restricted to "country" 826 and "subnatio" 2 then 4,120 observations remain. This is the same number as identified by Gill (2002 page xii bottom paragraph) in the national report.

The "country", "Subnation" and "stratum" variables can be used in PISA 2003 to identify children from England. If the data are restricted to "country" 826, subnation 8261 and "stratum" 82611 then 3,766 observations remain. This is the same number as identified by Micklewright and Schnepf (2006 page 13, third paragraph from the bottom).

The "country", "Subnation" and "stratum" variables can be used in PISA 2006 to identify children from England. If the data are restricted to "country" 826, subnation 82610 and "stratum" 82611 then 4,935 observations remain. This is the same number as identified by Bradshaw et al 2007 page vii point 2.1) in the national report.

Change of test month

In the national report for PISA 2000 Gill et al (2002 page 140 last paragraph) state that:

"Testing took place at the earliest opportunity in England, in March to mid-May 2000" [my emphasis]

On the other hand, in the PISA 2006 national report Bradshaw et al (2007a page 1 fourth paragraph) note:

"In England, Wales and Northern Ireland, students sat the two-hour assessment in November 2006 under test conditions" [my emphasis]

as does Bradshaw et al (2010 page 1 fourth paragraph) regarding PISA 2009:

"In England, Wales and Northern Ireland, pupils sat the two-hour assessment in November 2009 under test conditions" [my emphasis]

The month when the assessment was conducted has clearly changed. But what was the reason for this? Firstly, note that the UK (as a whole) was excluded from the PISA 2003 international report because of the low response rates (Micklewright et al 2006 page iii third paragraph) and concerns that the data would therefore suffer from response bias. This problem obviously needed to be addressed in future waves. Further insight comes from Bradshaw et al (2010 page 7 second complete paragraph) who state:

"Countries were required to carry out the survey during a six-week period between March and August 2009. However, England, Wales and Northern Ireland were permitted to test outside this period because of the problems for schools caused by the overlap with the GCSE preparation and examination period. [emphasis my own]*"*

This implies that low response rates for England in PISA 2000 and 2003 may have been influenced by the close proximity to children's GCSE exams (the PISA test was conducted just 2 months beforehand). To try to limit the burden of PISA on participating schools in 2006 and 2009 the test was moved to much earlier in the academic year.

Change from age to grade based sample

As noted in the main body of the paper, PISA is based upon children born within a calendar year. The OECD set strict rules that the pupils tested should be between 15 years and 3 completed months and 16 years and 2 completed months at the beginning of the test period (variation of up to one month in this definition is allowed). As England was given special dispensation to conduct the PISA survey in November, the vast majority of pupils who met this age definition (4,065 out of the 4,081) were in year 11. Hence, although the 2006 and 2009 data for England does meet the OECD requirements of all children being between 15 years 3 months and 16 years 2months, they are (for all intents and purposes) grade based samples.

The international dataset (downloaded from the PISA website http://www.pisa.oecd.org) also includes information on children's month and year of birth and the school grade they are in. To illustrate how the change from an age to a grade based sample made a difference to the sample composition for England, Appendix Table 1 presents summary statistics based on these three variables.

Appendix Table 1

Notice that in PISA 2000 and 2003 children were all born in the same calendar year (e.g. 1984) but belonged to different school grades (year 10 or year 11). In particular, the September to December born children were all in year 10 (the penultimate year of compulsory schooling in England). However, in PISA 2006 and 2009 the opposite held true. Children were born in different calendar years (e.g. 1993 or 1994) but all belonged to the same school grade (year 11). This provides clear evidence that there has been a change to at least some of the PISA survey procedures between the different waves.

This would seem to be confirmed by the national reports for England. Gill (2002 page 21) state that in PISA 2000:

"Two-thirds of them [sampled pupils] *were in Year 11 and one-third were in Year 10"*

In contrast, Bradshaw et al (2007, page 14) describe the PISA sampling procedures in the 2006 wave:

> *"The schools which had been selected in the sample were then invited to participate, and those which agreed were asked to supply details of all students who would be in Year 11 at the time of the beginning of the PISA survey period in November 2006* [my own emphasis]*"*

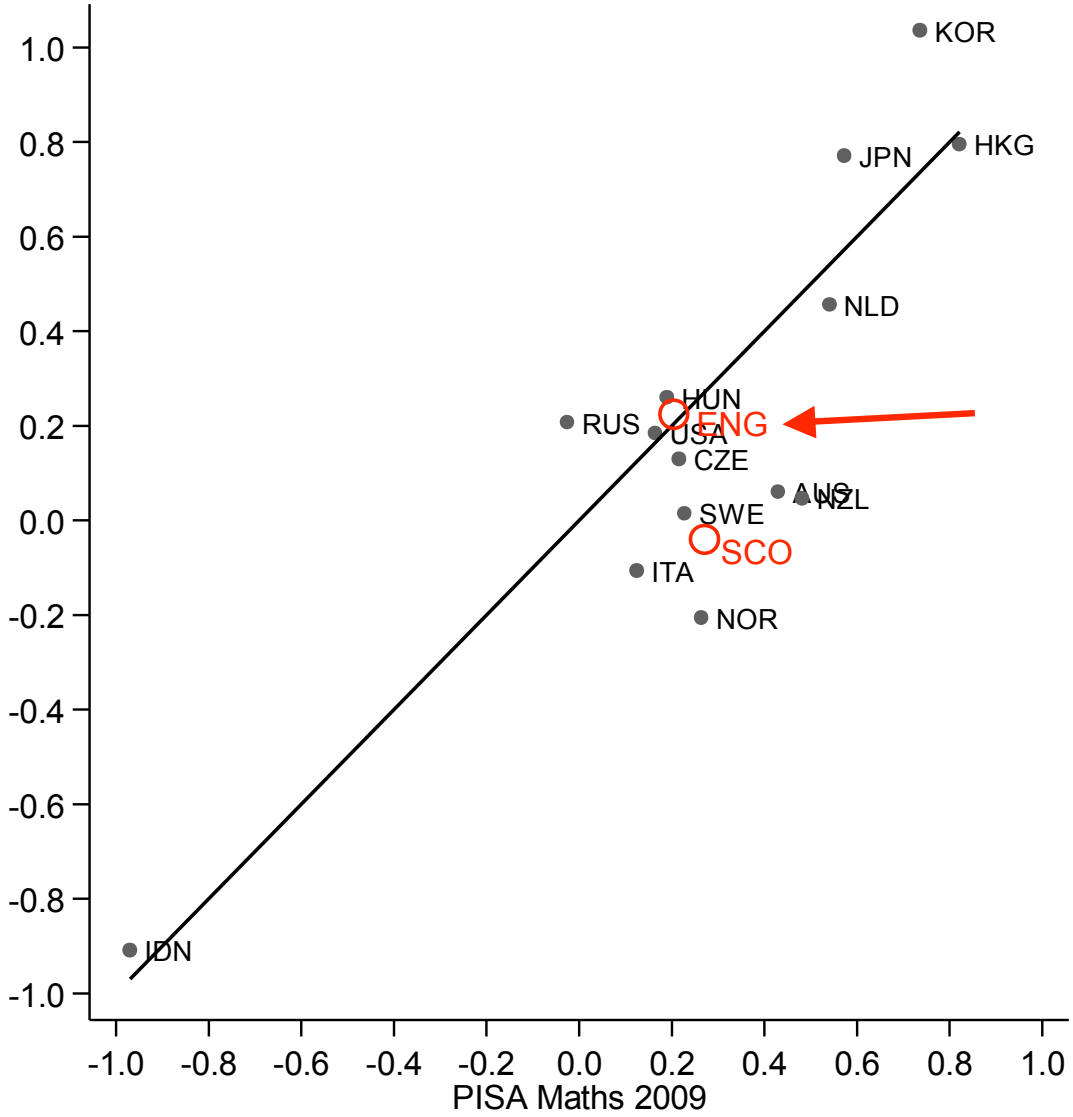
Appendix 5. The change in mean PISA and TIMSS test scores in Scotland

It is interesting to consider the results for England to those for Scotland. These two countries have obvious similarities (e.g. language and location) but also some important differences (e.g. aspects of the schooling system). Of particular interest with regards this paper is that the PISA data for Scotland does not seem to have at least some of the data problems that have been highlighted in the case of England. Specifically, the test in Scotland seems to have been conducted in the same month (March) across all PISA survey waves, with an age-based sample maintained throughout. The OECD also seems to have less concern with regards to non-response bias in this country, as it was included in their recent report on changes over time (see Annex B2 of OECD 2010a). Interestingly, PISA test scores for Scotland still show evidence of a significant decline. For instance, mean maths test scores dropped from 533 in PISA 2000 to 524 in 2003, 506 in 2006 and 499 in PISA 2009. What then should one take from such results? Although it is dangerous to generalise from Scotland to England, the findings in this Appendix perhaps to some extent support those who believe that there has been a genuine decline in secondary school children's performance over time.

Appendix Table 1. The birth month, birth year and "grade" of the 2000 – 2009 PISA England samples

	PISA 2000		PISA 2003		PISA 2006		PISA 2009	
	Birth year	Grade	Birth year	Grade	Birth year	Grade	Birth year	Grade
Birth month	st01q03	% Year 11	st02q03	% Year 11	ST03Q03	% Year 11	ST03Q03	% Year 11
January	1984	98.7	1987	82.3	1991	100.0	1994	99.8
February	1984	99.3	1987	82.0	1991	100.0	1994	100.0
March	1984	98.4	1987	99.4	1991	99.5	1994	99.7
April	1984	98.6	1987	99.0	1991	99.3	1994	99.7
May	1984	99.1	1987	98.8	1991	99.8	1994	99.4
June	1984	98.0	1987	98.7	1991	99.8	1994	99.7
July	1984	98.6	1987	99.0	1991	99.3	1994	99.7
August	1984	96.4	1987	99.7	1991	100.0	1994	98.9
September	1984	2.2	1987	0.9	1990	99.5	1993	99.4
October	1984	1.0	1987	0.6	1990	99.8	1993	100.0
November	1984	0.5	1987	2.0	1990	100.0	1993	99.7
December	1984	0.6	1987	0.3	1990	99.4	1993	99.4

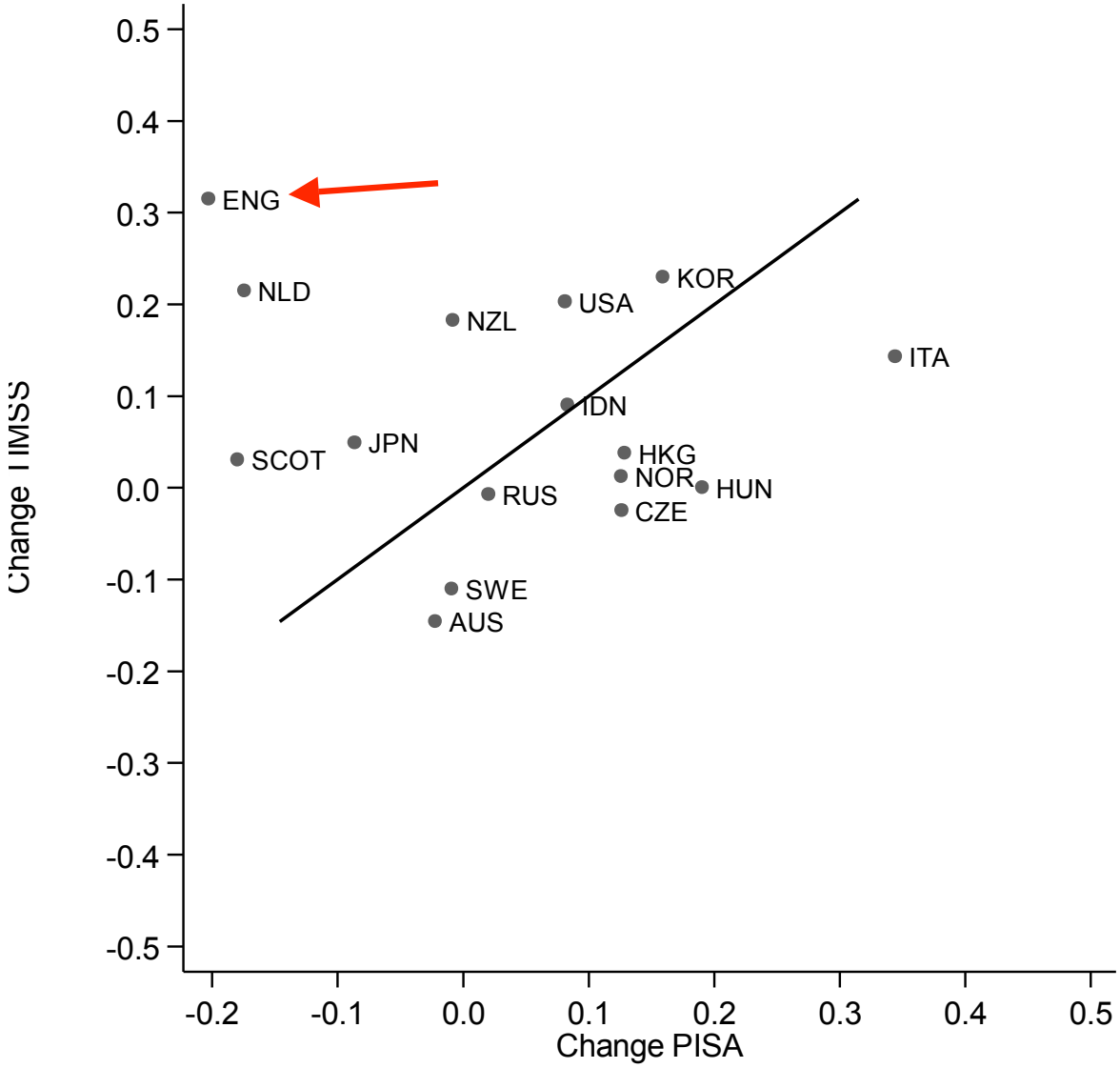
Appendix Figure 1. Results for TIMSS 2007 and PISA 2009 maths scores for a pool of 16 countries



Notes:

Figures are presented in terms of international z-scores, with the data having standardised within the sub-set of the countries considered. PISA 2009 maths test scores sit on the x-axis while TIMSS 2007 scores run along the y-axis. The solid 45 degree line represents where mean test scores in PISA are the same as those for TIMSS. Data for the Netherlands and New Zealand refer to TIMSS 2003 as these countries did not take part in the 2007 wave.

Appendix Figure 2. Change in PISA maths scores 2000-2009 compared to TIMSS maths scores 1999 – 2007 for a pool of 16 countries



Notes:

For most countries figures running along the x-axis refer to the change in mean PISA test scores between 2000 and 2009, while those along the y-axis are in reference to the change in TIMSS scores between 1999 and 2007. TIMSS data from New Zealand and Netherlands refer to 1995 to 2003 comparison, while for Norway, Scotland and Sweden it refers to the 1995 to 2007 comparison. All figures presented are in terms of international z-scores. The solid black line represents where the change in PISA test scores over the period is the same as the change in TIMSS test scores. In the Netherlands the total participation rate in TIMSS 1995 was 23% before replacement and 60% after replacement. In TIMSS 2003 this increased to 79% and 86% respectively.