Standards in Education: Reforms, Stagnation and the Need to Rethink

David Bolden[a*] and Peter Tymms[a]
[a]School of Education, Durham University, Durham, UK, DH1 1 TA

Abstract

Countries around the world are striving to improve their educational systems with a view to improving their economy and society. In this global competition national and international test results are of considerable interest. In this paper, we show that national testing in England and the USA have shown little or no improvement over the years. This finding is not isolated; it appears to be a global phenomenon. Data from large-scale international assessments such as PISA, TIMSS and PIRLS are remarkably stable over time. This paper reviews the trends from country-specific and international data and explores some of the reasons which have been offered for such stability. We argue that these explanations are insufficient and ways forward are discussed.

Keywords: standards, reform, assessment, TIMSS, PISA, PIRLS

Background

Countries around the world are involved in huge efforts to raise educational standards. This is reflected in the existence and growth of national and international bodies devoted to the measurement of educational performance. As an example, consider Cambridge Assessment which had a revenue of £129 million in 2002, but, by 2018 this had risen to £413 million and the proportion of revenue coming from overseas had risen from 25% to more than 80% (Lebus, 2018). Nations often have their own assessment bodies but large-scale international organisations are also important. They include the Organisation for Economic Co-operation and Development (OECD) which runs the Programme of International Student Assessment (PISA) (http://www.oecd.org/pisa/), and the International Association for the Evaluation of

* Corresponding author. David Bolden, Durham University, School of Education, Leazes Road, Durham, DH1 1TA. Email: d.s.bolden@durham.ac.uk.

Educational Achievement (IEA) which runs the Trends in Mathematics and Science Study (TIMSS) (https://www.iea.nl/timss) as well as the Progress in International Literacy Study (PIRLS) (https://nces.ed.gov/surveys/pirls/) are also testimony to the trend mentioned above.

Over the last thirty years, huge efforts have been made by successive governments, in England, to raise educational attainment and the education system has experienced massive reform, at huge financial cost. The National Curriculum was set up in 1988 by an Act of Parliament and Government-funded agencies were created (and some subsequently disbanded) to monitor the effects of these reforms (Whetton, 2009). This was achieved by tests of pupil performance in key areas at the ages of 7, 11 and 14 years (Dearing, 1993) and through the introduction of a harsh inspection system (Ouston, Earley, & Fidler, 2017). The cost has not, to our knowledge, been totalled, including as it does, private consultants, private tutoring and the National Strategies amongst many other things. However, we know that the testing of 11 year olds, at the end of 2007 alone, cost £18.9 million (Tymms & Merrell, 2009) and, the cost for Ofsted in the 2005/06 financial year was just under £220 million (Ofsted, 2006). These reforms might be thought to split neatly into those that were designed to improve standards and those that were designed to monitor impact. However, the testing and inspection system can be seen as part of the improvement mechanism; they aimed to hold schools to account. The domestic reforms, outlined above, are not generated in a vacuum; they are strongly influenced by global forces such as the OECD, the World Bank and the World Trade Organisation (Moutsios, 2009; Niemann, Martens, & Teltemann, 2017).

The United States, too, has been influenced by global movements and has seen major educational reform over a similar time in an effort to raise standards. A recent OECD report shows that the US spends more than most other participating countries on the education of its pupils. This amounted to an average of $16,268 per student per annum (from primary/elementary to tertiary education) compared to the OECD average of $10,759

(OECD, 2017). A 2011 news article reported that federal spending on education had increased by nearly 64% in the years after the implementation of its most recent significant reform, the 2001 No Child Left Behind Act (CNS News, 2011). The US have also been more assiduous than other countries in tracking standards over time. The National Assessment of Educational Progress (NAEP) has two main components. The main component collects data on children's performance in English and mathematics at Grade 4 and Grade 8 (children typically aged 9 and 14 years of age respectively) and began in the 1990s. These tests change with each cycle to reflect changes in the curricula operating at the time. The second component is, what is termed the long-term trend (LTT), and this has collected data on the reading and mathematics performance of 9-, 13- and 17-year old children in public and private schools since the early 1970s; it purports to assess the same knowledge and skills (see for example Vanneman, Hamilton, Anderson, & Rahman, 2009).

A number of academics (e.g. Slavin, 2002; Turner et al., 2003) have questioned the warrants of national education policies. These criticisms are sometimes accompanied by a call for more rigorous, scientifically based, evidence to lead to so-called evidence-based policymaking. Partly in response to this debate, many countries are now also spending significant sums on large-scale randomised control trials (RCTs) as a way of evaluating educational interventions; the Education Endowment Foundation (EEF) in England, the National Center for Educational Evaluation and Regional Assistance (NCEE) in the US, the European Schoolnet in Europe, the Social Ventures in Australia, the Jacob's Foundation in Switzerland, the Nippon Foundation in Japan and the Lemann Foundation in Brazil have all moved towards using large-scale RCTs (see Lortie-Forgues & Inglis, 2019).

RCT methodology is not without its critics, such as Hammersley (2005). However, he has more recently clarified his position writing: "*RCTs can be a very useful method. …. Their*

distinctive value lies in the fact that random allocation to groups receiving different 'treatments' greatly reduces the danger of selection bias … In this specific respect, RCTs are superior to non- experimental quantitative research and to qualitative work." (Hammersley, 2015)

Recent results from some of the efforts to discover 'what works' will be used to throw light on the potential of RCT evidence to raise standards i.e. levels of attainment.

This paper aims to explore the results from a number of large-scale assessments and show that there is little evidence of the impact of educational reforms. We then discuss some possible reasons for the lack of progress and tentative ways forward.

The Data Sources

We first review data from England before moving on to the USA and then international data. We describe the major efforts made by these countries in terms of the educational reforms implemented and describe the effects of these on the educational outcomes of their populations.

From England

Since the implementation in 1988 of the Education Reform Act, all state-funded schools in England have had a centrally designed national curriculum. This curriculum prescribed the content that all state-funded schools were required to teach within each subject domain. Introduced alongside the new curriculum was a programme of national testing. The curriculum was divided into four Key Stages and at the end of Key Stages 1, 2, and 3 at ages 7, 11, and 14 respectively, all eligible children were tested in key subjects (mathematics, English and science). The results of these tests were intended to hold schools to account and were presented as the percentage of children who had reached each of the 'expected levels' within each Key Stage.

We examine national data from the end of Key Stage 2 for English and mathematics where the key indicator is the percentage of children reaching the so-called Level 4. The data from each publicly funded school are published by the Department for Education and appeared in league tables. Figure 1 below shows the trend in the percentage of children reaching level 4 in mathematics and English between 1995 and 2015[1] from full national cohorts.

[Figure 1 somewhere here]

The data for both English and mathematics show a dramatic rise between 1995 and 2000, but then, from 2000 onwards there was a 'levelling out' of the trend with only small variations thereafter.

Not surprisingly, the English government at the time used the steep rise from 1995 to 2000 to justify and promote the efficacy of their educational reforms (see for example the paper by the key educational advisor at the time; Barber, 2001). However, data from other sources suggested that this steep rise in attainment was misleading. For example, Tymms (2004) questioned the discontinuity between the steep rise we see from 1995 to 2000 and the trend thereafter. He used a large amount of independent performance data to show that this steep rise pre-2000 was, in the words of Massey, Green, Dexter, & Hammet (2003), 'illusory'. The analysis by Tymms showed that what actually happened was a much more modest increase between those two time points. The rise amounts to an effect size of about 0.25 (Cohen's d) and the mathematics rise to about 0.55; an order of magnitude which might result from test practice. After 2000, the results for both English and mathematics have risen by small

[1] We only show data up to and including 2015 for mathematics because a new curriculum was introduced in 2014 and testing arrangements changed from 2016. Data for English is only shown up to and including 2011 as new tests of English began to change from 2012 onwards and therefore data from that time are not directly comparable to previous years (e.g. new grammar, punctuation and spelling tests were also introduced in 2013).

amounts, but since there are no independent data, that we are aware of, we cannot be sure that standards (levels of attainment) have risen. However, the information shown in Figure 1 is suspicious. The mathematics and reading lines surely hug one another too closely; any policy impact would be more effective for one of them. We might expect both to rise, but not that the differences between the percentages would be almost identical in 1995 and 2011.

The pattern seen in England follows the well-known phenomenon that, whenever tests are introduced, or changed, on a large-scale, scores tend to rise (see Koretz, 2011; Klein, Hamilton, McCaffrey, & Stecher, 2000) even when there is no increase in levels of attainment; so well-known is the effect it that has been given the name 'Campbell's Law' which parallels 'Goodhart's Law' (Rodamar, 2018). A very recent analysis of the phenomenon by Cuff, Meadows, and Black (2019) identifies 'test familiarity' as the underlying cause of the rise in test scores. So, the general pattern we see in Figure 1 is of little surprise to educational researchers with an international perspective. There were, however, additional factors at play here and which helped produce this 'illusory' rise. Tymms (2004) indicated that the pre-2000 rise was partly due to a faulty standard setting procedure. The fault was that the cut-scores, for deciding which students were awarded which levels, were made to equate each year with the immediately preceding one. A slight leniency each year resulted in a substantial rise over several years. The problem is akin to the errors in copying made by scribes reproducing medieval manuscripts from faulty copies. Once the problem was identified in 2000, the steep rise plateaued.

There were also significant increases in the pass rates for the national tests at the end of compulsory schooling at age 16 years (GCSE) and at the pre-university level at age 18 years ('A' level). But the use of independent anchor tests have shown that these rises, up to 2006, are, most likely, the result of grade inflation (Coe, 2007; Coe & Tymms, 2008).

In summary, the data suggest that, despite the very many educational reforms introduced in England over the last 30 years, and the many millions of pounds in associated costs, the results show very little substantive change in academic levels.

The US data

The United States too has seen major investment over the last 30-40 years in education, and the implementation of substantial reforms in efforts to raise educational standards. Figures 2 and 3 below show the long term trend (LTT) in reading and mathematics performance. As noted earlier since the 1970s, NAEP has collected data with the aim of generating a consistent body of knowledge and skills over a long period using equivalent tests over the years.

The average reading and mathematics results show some slight improvements for 9- and 13-year olds between 1971 and 2012 (Figures 2 and 3), but almost no change for 17-year olds. Further, the general trends for all age groups for both reading and mathematics, and the spread of scores (standard deviations), are remarkably stable given the lengthy time period over which reforms have been implemented.

[Figure 2 somewhere here]

[Figure 3 somewhere here]

The No Child Left Behind (NCLB) Act (2001) made it mandatory for all US states to design accountability systems and to conduct annual assessments of all public school students with

the aim of identifying schools that were failing to make adequate yearly progress in reading and mathematics. If the NCLB agenda had had some impact on children's attainment in mathematics and reading then this would surely be reflected in the results from the NAEP data post-2001.

Dee & Jacob (2011) analysed NAEP data from state-level test scores pre- and post-NCLB reform and reported some modest improvements in mathematics achievement (ES = 0.23) by 2007 but no impact in reading. Fuller, Wright, Gesicki, & Kang (2007), again using NAEP data on fourth graders, concluded that the change in attainment has actually become flatter since the introduction of NCLB. Similarly, an analysis of the NAEP reading by Lee (2006) found it to be stable over the NCLB period. Others have suggested that the high stakes nature of the testing regime inherent in the NCLB agenda creates 'incentives' for teachers to teach to the test and this can lead to misleading conclusions. For example, after an analysis of students' performance on both the English and mathematics tests in three states, Jennings and Bearak (2014) showed that students performed better on items that were tested frequently, suggesting that the results may have over-stated students' performance.

International Data

This section reviews the data from the major large-scale assessments specifically PISA, TIMSS and PIRLS.

PISA Data

PISA tests of reading literacy, mathematics literacy and science literacy are taken on a three-year cycle by a representative sample of 15-year old students from each participating country. It has been running since 2000 with OCED and non-OECD countries and, since its start, the numbers of countries and students have increased steadily (see: https://nces.ed.gov/surveys/pisa/countries.asp). During each cycle, not all students are

exposed to the full range of items for each literacy. Rather scores are imputed using the available data. It is also worth noting that PISA aims to report tests on the same scale at each time point but that some items in the test are changed with some being released to the public and others remaining undisclosed. The aim is to maintain a common standard for the assessments but to allow others to see examples of test content.

PISA results have remained remarkably consistent with correlations between each country's mean scores across time points being extremely high, ranging from 0.89 to 0.99 (see Aloisi & Tymms, 2017). Their detailed analysis showed that reforms in PISA-participating countries show little or no impact on their respective mean scores in reading and mathematics.

There is a tendency for the scores to rise across successive testing cycles. But the OECD has acknowledged that student populations and economic success are liable to change and that this may have an impact irrespective of educational policy. The changes include factors such as age, SES, migration status and first language. Consequently, the OECD have started to produce 'adjusted' trends that attempt to control for such factors. Recent work by Aloisi and Tymms (2017) show the importance of controlling "non-policy-malleable variables" of the student cohorts. They write that "… *there is strong evidence showing that SES, immigration status, and grade are related to PISA outcomes …. [and that] the picture that emerges from OECD analyses shows that the difference between the usual reported scores and the adjusted ones is not negligible*" (p. 184-5).

However, a number of studies have suggested that policy reforms have had an impact of international test results; how do they square with the Aloisi and Tymms' (2017) claims? Specific high profile cases are considered.

Indonesia undertook significant reforms in 2003, introducing more stringent teacher certification as well as a new competency-based curriculum. Despite an increase in their

PISA scores in 2006 (see Barrera-Osorio, Garcia-Mareno, Patrinos, & Porta, 2011) there was an overall flat trend between 2003 and 2013 once socio-demographic changes are accounted for (OECD, 2014). Further support for this lack of impact comes from Indonesia's rather static scores in TIMSS over the same period. Others have argued that any change in Indonesia's performance in reading and mathematics was the result of factors outwith policy changes, i.e. changes in the socioeconomic profile of the PISA cohorts (Aloisi & Tymms, 2017).

Data from policy reforms in Qatar in 2001 produced some promising improvements in both PISA and TIMSS scores over the decade that followed, but other evidence suggests the improvements were largely not the result of these policy changes. For example, research by Areepattamannil, Melkonian, & Khine (2015) indicate that the main cause behind the improvement in PISA mathematics scores probably lies in the changing demographics during that time. More specifically, Qatar was experiencing significant immigration during that period and such students have been shown to routinely outperform native learners. This suspicion is further strengthened when one looks at PISA scores once they have been adjusted for socioeconomic and demographic change; the improvement between 2009 and 2012 disappears (OECD, 2014).

Qatar is one of the six Gulf monarchies studied by Mohammed and Morris (2019). They note how the major educational reforms were generated by 'policy borrowing' in a very specific sense. It "…*involves the GEI [Global Education Industry] selling a range of products that are described as 'best global practices'. These are portrayed as the transferrable sources of success in education systems that performed well on international tests*." (p. 13). Their conclusion about the impact of the expensive reforms which started in 2001 is that the Gulf approach "… *appears to have failed to achieve its specified outcomes and created a cycle of dependency*" (p.16).

We can find similar patterns in other countries where the impact of policy reforms seems to have positive results but which can be equally or better accounted for by other, non-policy malleable factors e.g. Ireland and Belgium. Overall, Aloisi and Tymms (2017) conclude from their analysis that there is "*no strong evidence for the effectiveness of curricular reforms …*" (p. 205).

TIMSS

TIMSS has been conducted on a four-year cycle since 1995 and tests children in Grade 4 (typically aged 9-years) and Grade 8 (typically aged 14-years) in mathematics and science. The survey was last conducted in 2015. TIMSS is designed to align broadly with mathematics and science curricula in the participating countries and is used to compare knowledge and skills of participating students over time in those areas. Participation has grown over consecutive cycles and in 2015 TIMSS was administered in 49 IEA member countries and 6 other education systems at Grade 4, and in 38 IEA member countries and 6 other education systems at Grade 8. Performance is TIMSS is measured using a scale average of 500 and a standard deviation of 100 which are designed to remain a fixed point of reference across the different cycles.

The most recent analysis of TIMSS states that "*both long term and short term trends are up in both subjects and at both grades and that more countries registered increases than decreases from 1995 to 2015 and from 2011 to 2015*" (see TIMSS, 2015). At a finer grained analysis, we find that in mathematics between 2011 and 2015, 23 countries increased their achievement scores whilst 5 showed a decrease and 15 remained the same.

To illustrate the trend in TIMSS scores we have taken the average results from those countries who have thus far participated in all cycles (See Figs, 4 and 5 below). This list changes slightly depending on whether the focus is Grade 4 or Grade 8.

[Figure 4 somewhere here]

[Figure 5 somewhere here]

On average the scores have risen. The gains between 1995 and 2015 equate to Effect Sizes (ESs) of 0.15 and 0.23, for maths in grade 4 & 8 respectively, and 0.17 and 0.23 for science in Grades 4 & 8. This amounts to and ES of about 0.01 per year.

As with PISA data mentioned earlier it could be that small gains in the TIMSS scores can be accounted for by 'non-policy-malleable variables' of the student cohorts such as SES. For example, in Turkey, Atar & Atar (2012) examined the effect of recent educational reforms on Turkish students' science scores in TIMSS and found that these increased in parallel with an increase in the SES of their families. On the other hand, it could be that extra time allocated to specific subjects could explain the gains.

PIRLS

The Progress in International Reading Literacy Study (PIRLS) began in 2001 and since its inception, it has tested 10-year-old students from participating OECD countries on a five-year cycle. Each cycle has typically seen an increase in the number of participating educational

systems[2], from 36 in 2001 to 61 in the last cycle of 2016. PIRLS is also measured using a fixed scale.

[Figure 6 somewhere here]

We find a similar story with the PIRLS data as we did with TIMSS; a slightly rising trend in average scores for those countries participating in all cycles but the rise is small (ES = 0.11). It amounts to an ES of less than 0.01 per year (see Fig 6).

Some Possible Explanations

In the sections above, we have tried to give the reader some insight into the huge amount of effort and money invested by countries all over the world into raising educational standards. We have also reviewed data, which suggests that educational policy changes have had little impact in developed countries over the last thirty years. This leads to a fundamental question: how much improvement would we expect from the reforms? To our knowledge, no serious academic has claimed that such and such an intervention, or combination of policies, on a national basis, will increase maths, or other attainment, by a specific amount. Of course, there have been claims by politicians about 'transformation' (e.g. Blair, 2004) and national policies leading to 'world-class' education (e.g. Morgan, 2016). But, such political rhetoric is just that. The absence of quantified claims suggest a global lack of deep understanding of educational systems. Nevertheless, it is clear that efforts at transformation have met with failure on a massive scale. Why is that? We outline possible reasons, which brings together the work of a many academics. They are meant to be illustrative of kinds of issues that must

[2] The term 'educational systems' is used to denote the fact that some are countries whilst some are subnational entities, e.g. provinces of Canada, etc.

be taken into account when trying to understand what has happened; they are not exhaustive and they are set out in two categories – Implementation and Omission.

1. **Implementation**: things that were done but should have involved more thought:
    a. **Timing**
        i. **Speed of reform**: The first explanation stems from the observation that governments change often and reform quickly. Most national governments are typically in office for only 4 to 5 years at a time before having to seek re-election. This means they need to show their electorate very quickly that they are doing a good job and there can be a tendency to rush through reforms without the necessary research base. One example comes from Higgins (2012), it involved an evaluation of the impact of the introduction of interactive whiteboards in schools in the UK. Higgins writes "*In terms of the politics of the evaluation, the analysis of the final results came after the decision had been made to expand the pilot, which was based on (or bolstered by) the interaction and perception of data available earlier in the evaluation*" (p. 135). The result was the roll out of a policy that was not backed by evidence.
        ii. **Frequency of reform**: Another related explanation comes from the frequency with which policy initiatives are introduced. As an example, Tymms (2001, p. 3) notes that reforms in England have included *"… the introduction of a National Curriculum, national testing, a heavy inspection regime and hundreds of lesser reforms*". It is not too far-fetched then to imagine that in such a climate some teachers may not fully embrace the change because they know another one will be along soon. In addition, when there is a flurry of reforms the positive impact of some may be

cancelled by the negative impact of others. In 1997 in England, David Blunkett was appointed as Secretary of State for Education and Employment in the newly elected Labour government. During the first four years of his tenure he introduced 60 (sic) initiatives related to education (Guardian, 2001).

At this point, the reader might expect an analysis relating 'between country variability' to 'gains in international tests'. However, as was noted earlier, the international test data are so stable that no such analysis could yield clear relationships. Neither short-term reforms, such as the introduction of whiteboards, nor the major long-term changes from Indonesia made a difference. It seems that the speed of reform may be problematic but it is not a complete explanation of the failure of reforms.

Similar points can be made about the frequency of reform. Given the frequency of reform and the stability of international tests we, as international researchers, are not in a position to compare countries' changing tests scores.

b. **Carrots and sticks**: Reforms are often accompanied by measures designed to ensure compliance, such as tests, league tables of test results, high-stakes inspections and financial penalties/rewards. These measures/incentives can result in unintended consequences (see for example Smith, 1995; Fitz-Gibbon, 1997; and Jones et al., 2017) which can negate the original intention of the reforms. One might still argue that the unintended negative consequences are outweighed by the intended positive impacts. However, the whole gamut of reforms in England (curriculum specification, testing, inspection etc.) were designed to hold

schools to account, but no change in test scores were observed. Similar points can be made about the USA as well as Indonesia and Qatar. It seems that such reforms, in the form that they have been implemented have not worked.

c. **Posturing**: A long line of policy analysis suggests that policy, or at least the rhetoric surrounding policy, has more to do with posturing that it does with raising attainment or other important matters (see for example Adamson et al., 2017; Crossley, 2019).

2. **Omission:** what has not been taken sufficiently into account?

a. **Factors outside school**: It has long been recognised that student success is not just a product of schooling but is related to factors outwith the formal educational system. Bourdieu is probably the most cited and influential theorist in this area. He argued that students acquire a degree of Cultural Capital from their background and that this is transferred from generation to generation maintaining a social structure across the years. Cultural capital is seen as "*linguistic and cultural competence and that relationship of familiarity with culture which can only be produced by family upbringing when it transmits the dominant culture*" (1977, p. 194).

Those working in the Comparative Education tradition have also noted the importance of non-school factors. For example, Deng and Gopinathan (2016, p. 461) write as follows: "*Explaining the PISA success of an Asian country like Singapore is by no mean easy; there are many potential contributing factors – Confucian cultural orientation, students' motivation, resilience, time spent on homework, modes of learning and out-of-school tuition, among others.*"

Academics in the School Effectiveness tradition have also noted the importance of

non-school factors. For example, Tymms, Merrell and Wildy (2015) developed a theory which holds "*that variables which are closest to the student are the most influential but that the jurisdiction where the student is educated, which has its own approaches to education and upbringing is of similar importance*" (p. 356). Within the same tradition a technique known as regression discontinuity analysis has been used to estimate the amount of learning which, shown by test results over a specific period, can be ascribed to schools and which to non-school factors (Luyten, 2006). On balance, although the ratio varies across countries, schooling appears to be a little more important than non-school factors.

b. **Changing pedagogy:** There is now a solid body of evidence to suggest that efforts at changing teacher practice is very difficult and changing teacher practice such that it has a positive impact on student outcomes is even more complex and difficult to achieve (see for example Timperley, Wilson, Barrar, & Fung, 2007; Adey, Hewitt, Hewitt, & Landau, 2004). This implies that structural changes to curriculum and reward structures are unlikely to result in meaningful long term change.

c. **Transplantation**: The idea of 'transplanting' good practice from one context to another (whether that be from top performing countries or schools) is common but questionable (Scheerens, Luyten, Van den Berg, & Glas, 2015). One explanation of why the results are not always as expected comes from the idea of the 'loose coupling' of educational systems (Weick, 1976), i.e. "*shared expectations do not automatically mean shared action or implementation of these expectations at either the macro or micro levels*" (Wiseman & Chase-Mayoral, 2014, p. 107). A not dissimilar explanation is offered by Schweisfurth & Elliott (2019) and Elliott, Stankov, Lee, & Beckmann (2019). The former suggest that the introduction of

perceived ‘best practice’ from one country to another is likely to meet resistance, chiefly because “… *indigenous day-to-day practice is embedded in its context, underpinned by societal expectations and norms which reflect the complex workings of culture”* (op cit: p.1). The latter argue that large-scale assessments can make a useful contribution to knowledge but need to be seen as one factor in many.

d. **Resistance:** A related explanation concerns the resistance that people exhibit to changing their views. A thorough account of the literature comes from Chin and Brewer (1993) in which they show how information which challenges beliefs (anomalous data) is readily discounted. Various mechanisms are identified and change can only be brought about with great effort. They argue that *“understanding how people respond to anomalous data is crucial to understanding the process of theory”* (p 39). One wonders how many policies reforms have come to grief because the reformers have simply not got to grips with the mechanisms needed to change minds and behaviour.

e. **Nature and nurture**: A fundamental reason why educational reforms have not impacted on attainment levels might be related to the role of the environment, including the home, and biological factors involved in academic progress. We break this section into two parts: genetics and other non-educational influences in cognitive development

    i. **Genetics:** consider these quotes from Plomin (2018) linking genetic makeup to educational success: “*Mothers matter but they don’t make a difference*” (p. 167). “*Polygenetic scores are also the best predictors of how well children will do at school*” (p. 160). And *“...20 percent of the variance in school achievement could be due to school or home*

environments, although this effect mostly washes out by the time children go to university" (p. 95). Such statements provide the basis for much debate and we are mindful of the dangers of taking an erroneously determinist position. Ridley (2003) makes it clear that genes do not work in isolation and that it is through interaction with our environment that we become what we are. We also need to be aware that heritability data relate to populations as a whole and should not be used to draw conclusions about individuals. However, there can be little doubt that educational systems are not working with blank slates (Pinker, 2004). Pupil-level factors are important predictors of educational outcomes (Creemers & Kyriakides, 2015; Scheerens et al., 2015).

ii. **Developmental:** A number of non-educational and non-genetic factors have a lasting detrimental impact on children's cognitive development. We mention four to illustrate the importance of these largely unmentioned (in the educational literature) features, which must surely have a bearing on educational attainment across countries. The first is foetal alcohol syndrome; if the pregnant mother drinks too much alcohol, the unborn child's cognitive development can be permanently impaired (Guerri, Bazinet, & Riley, 2009). The second is exposure to lead. It has long been known that this can have a harmful effect but the issue remains (Reuben 2017). The third is domestic violence, which is universally condemned, but the impact on a child's IQ is less well known (Koenen, Moffitt, Caspi, Taylor, & Purcell, 2003). Finally, we now know that chronic or acute stress of the mother during pregnancy can have a long-term negative impact on the unborn child (Baibazarova et al., 2013).

Conclusion

This paper has made it clear that educational standards, by which we mean, levels of academic attainment, of affluent countries, are incredibly stable over a decade and more. Further, major efforts at improvement have noticeably failed. From this, one lesson is clear; we need to alter the way we view large scale educational change. That is, we should think of national educational change as a process of small incremental improvements, which may accumulate over long periods of time (decades). We should expect these increments to be difficult to both attain and maintain. We should be sceptical of quick answers, which we can see, with hindsight, are often superficial and glib. If this could be widely accepted, it should deal with the first two explanations of the lack of change listed above: Speed and Frequency of Reform

Our analysis has implications both for policy making and for research. For policy making there are two broad ways forward. The first is a need to restructure the mechanisms of national policy making so that it is not closely tied to the short official lives of government ministers. Government needs to set a general direction but educational decisions relating to such matters as curriculum content, and national testing should be devolved to a non-political body which is set up for the long haul. The second, despite the cautionary words of Auld and Morris (2016), is the need for a permanent advisory research-based unit which can be called upon to provide evidenced-based advice on on-going issues. Such moves have the potential to address, if not solve, four of the explanations for the lack of change: Carrots and Sticks, Changing Pedagogy, Transplanting and Resistance. The importance of Biology can be highlighted by including medics in the advisory research-based unit whilst the insights of EEF, and similar bodies, can be similarly included.

For research the implications are profound.

We, as an educational research community, are suffering from collective ignorance: There is an enormous quantity of educational knowledge but, paradoxically, the most serious potential explanation for the general failure of educational reforms is our general lack of deep understanding of educational systems. Our present state of knowledge is simply not sufficient to advise policy makers about the impact of interventions. There is understanding but it is compartmentalised within a whole range of disciplines including psychology, sociology, genetics, medicine, economics, statistics and education. These disciplines do not communicate well with one another, if at all, and when they try they often find themselves talking at cross purposes. Even within disciplines, there is fragmentation. If we are to make progress, we need to take understandings on board from disparate groups and make sense of them with an overarching theoretical structure. Such a structure would be able to make testable predictions and be refined in the light of such tests. In calling for such a theoretical structure we realise that we are asking for something which is daunting, large scale and multi-disciplinary. The task is so large that it may be beyond us but the price of failure is to condemn educational policy making to repeated expensive failures.

References

Adamson, B., Forestier, K., Morris, P., & Han, C. (2017). PISA, policymaking and political pantomime: education policy referencing between England and Hong Kong. *Comparative Education*, *53*(2), 192-208.

Adey, P., Hewitt, G., Hewitt, J. & Landau, N. (2004). *The professional development of teachers: Practice and theory*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Aloisi, C., & Tymms, P. (2017). PISA trends, social changes, and education reforms. *Educational Research and Evaluation*, *23*(5-6), 180-220.

Andon, A., Thompson, C.G., & Becker, B.J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large Scale Assessments in Education*, *2*(1), 7.

Areepattamannil, S., Melkonian, M., & Khine, M.S. (2015). International note: Exploring differences in native and immigrant adolescents' mathematics achievement and dispositions towards mathematics in Qatar. *Journal of Adolescence*, *40*(April 2015), 11-13.

Atar, H. Y., & Atar, B. (2012). Examining the effects of Turkish education reform on students' TIMSS 2007 science achievements. *Educational Sciences: Theory and Practice*, *12*(4), 2632-2636.

Auld, E., & Morris, P. (2016). PISA, policy and persuasion: Translating complex conditions into education 'best practice'. *Comparative Education*, *52*(2), 202-229.

Baibazarova, E., van de Beek, C., Cohen-Kettenis, P. T., Buitelaar, J., Shelton, K. H., & van Goozen, S. H. (2013). Influence of prenatal maternal stress, maternal plasma cortisol and cortisol in the amniotic fluid on birth outcomes and child temperament at 3 months. *Psychoneuroendocrinology*, *38*(6), 907-915.

Barber, M. (2001). The very big picture. *School Effectiveness and School Improvement*, *12*(2), 213-228.

Barrera-Osorio, F., Garcia-Mareno, V., Patrinos, H.A., & Porta, E. (2011). Using the Oaxaca-Blinder decomposition technique to analyze learning outcomes changes over time: An application to Indonesia's results in PISA mathematics (Policy Research Working Paper No. 5584).

Blair, Tony. (2004). Speech https://www.theguardian.com/education/2004/jul/07/schools.uk3 (Retrieved 24/1/20).

Bourdieu, P. (1977). Cultural reproduction and social reproduction, in: J. Karabel, & Halsey, A.H. (Ed). *Power and Ideology in Education*. New York: Oxford University Press.

CNS News (2011). Available online at: https://www.cnsnews.com/news/article/education-spending-64-under-no-child-left-behind-test-scores-improve-little (Accessed 10th July 2019).

Coe, R. (2007). Changes in standards at GCSE and A-level: Evidence from ALIS and YELLIS: A report for the ONS (Durham, CEM Centre, Durham University).

Coe, R., & Tymms, P. (2008). Summary of research on changes in educational standards in the UK. Education Briefing Book 2008: IoD Policy Paper. M. Harris. London, Institute of Directors.

Crossley, M. (2019). Policy transfer, sustainable development and the contexts of education. *Compare: A Journal of Comparative and International Education*, *49*(2), 175-191.

Cuff, B. M., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 321-339.

Creemers, B., & Kyriakides, L. (2015). Developing, testing and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, *26*(1), 102-119.

Dearing, R. (1993). *The National Curriculum and its assessment*. London: School Curriculum and Assessment Authority.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418-446.

Deng, Z., & Gopinathan, S. (2016). PISA and high-performing education systems: Explaining Singapore's education success. *Comparative Education*, *52*(4), 449-472.

Elliott, J., Stankov, L., Lee, J., & Beckmann, J. F. (2019). What did PISA and TIMSS ever do for us? The potential of large scale datasets for understanding and improving educational practice. *Comparative Education*, *55*(1), 133-155.

Fitz-Gibbon, C. T. (1997). *The value added national project: Final report: Feasibility studies for a national system of value added indicators*. London: School Curriculum and Assessment Authority.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher*, *36*(5), 268-278.

Guardian (2001). Available online at: https://www.theguardian.com/uk/2001/mar/02/education.schools (Accessed 28th July 2019).

Guerri, C., Bazinet, A., & Riley, E. P. (2009). Foetal alcohol spectrum disorders and alterations in brain and behaviour. *Alcohol & Alcoholism*, *44*(2), 108-114.

Hammersley, M. (2005). Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice. *Evidence & Policy: A Journal of Research, Debate and Practice*, *1*(1), 85-100.

Hammersley, M. (2015, May). Was heißt hier eigentlich "Evidenz"?' Paper presented at Frühjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft für Evaluation (DeGEval), Fakultät für Sozialwissenschaften, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken.

Hess, F. M. (2011). *Spinning wheels: The politics of urban school reform*. Washington, DC: Brookings Institution Press.

Higgins, S. (2012). Evaluation of impact: A case study of the introduction of interactive whiteboards in schools in the UK, in: J. Arthur, M. Waring, R. Coe, & L. Hedges. (Ed). *Research Methods and Methodologies in Education*. London: Sage.

Jennings, J. L., & Bearak, J. M. (2014). 'Teaching to the test' in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381-389.

Jones, K. L., Tymms, P., Kemethofer, D., O'Hara, J., McNamara, G., Huber, S., & Greger, D. (2017). The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland. *Oxford Review of Education*, *43*(6), 805-822.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. & Stecher, B. M. (2000). What do test scores in Texas tell us? Santa Monica, CA: Rand.

Koenen, K. C., Moffitt, T. E., Caspi, A., Taylor, A., & Purcell, S. (2003). Domestic violence is associated with environmental suppression of IQ in young children. *Development and psychopathology*, *15*(2), 297-311.

Koretz, D. (2011). Lessons from Test-Based Education Reform in the U.S. *Zeitschrift für Erzhiehungswissenschaft, Special Issue 14*(Jan 2011), 9-23.

Lebus, S. (2018). Available online at: https://www.cambridgeassessment.org.uk/news/fifteen-years-of-change/ (Accessed 21st April 2019).

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An In-depth look into National and State reading and math outcome trends*. Cambridge, MA: The Civil Rights Project at Harvard University.

Lortie-Forgues, H., & Inglis, M. (2019). Most rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158-166.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, *32*(3), 397-429.

Massey, A., Green, S., Dexter, T. & Hammet, L. (2003). Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001. Final Report to QCA of the Comparability over Time Project (Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate).

Merrell, C., & Tymms, P. (2010). Changes in children's cognitive development at the start of school in England 2001-2008. *Oxford Review of Education iFirst Article*: 1-13.

Mohammed, M., & Morris, P. (2019). Buying, selling and outsourcing educational reform: The Global Education Industry and 'policy borrowing' in the Gulf. *Compare: A Journal of Comparative and International Education*, May 2019.

Morgan, Nicky. (2016). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/508421/DfE-strategy-narrative.pdf (retrieved 24th January 2020).

Moutsois, S. (2009). International organisations and transnational education policy. *Compare: A Journal of Comparative and International Education, 39*(4), 469-481.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Available online at: http://timssandpirls.bc.edu/pirls2016/international-results/ (Accessed 5th March 2019).

NAEP (2016). Available online at: https://nces.ed.gov/programs/coe/pdf/coe_cnj.pdf.

National Center for Education Statistics (2013). *The nation's report card: Trends in academic progress 2012* (NCES 2013-456). Washington, DC: Institute of Education Sciences, US. Department of Education.

Niemann, D., Martens, K., & Teltemann, J. (2017). PISA and its consequences: Shaping education policies through international comparisons. *European Journal of Education*, *52*(2), 175-183.

OECD (2014). *PISA 2012 results: What students know and can do (Volume 1, Revised edition): Student performance in mathematics, reading and science*, Paris: OECD Publishing.

OECD (2017). *Education at a Glance 2017: OECD Indicators*, Paris: OECD Publishing.

Ofsted (2006) Available online at: http://www.Ofsted.gov.uk/assets/Internet_Content/Publications_Team/File_attachments/Ofsted_Resource_Accounts05_06.pdf (Accessed 27th March 2007).

Ouston, J., Earley, P., & Fidler, B. (Eds.). (2017). *OFSTED inspections: the early experience*. London: Routledge.

Pinker, S. (2004). *The Blank Slate: The modern denial of human nature*. New York: Viking.

Popper, K. (1974). *Unended Quest.* London: Fontana.

Plomin, R. (2019). *Blueprint: How DNA makes us who we are*. Cambridge, MA: MIT Press.

Reuben, A., Caspi, A., Belsky, D. W., Broadbent, J., Harrington, H., Sugden, K., & Moffitt, T. E. (2017). Association of childhood blood lead levels with cognitive function and socioeconomic status at age 38 years and with IQ change and socioeconomic mobility between childhood and adulthood. *Journal of the American Medical Association*, *317*(12), 1244-1251.

Ridley, M. (2003). *Nature via nurture: Genes, experience, and what makes us human*. New York: HarperCollins.

Rodamar, J. (2018). There ought to be a law! Campbell versus Goodhart. *Significance*, *15*(6), 9-9.

Scheerens, J., Luyten, H., Van den Berg, S.M., & Glas, C.A.W. (2015). Exploration of direct and indirect associations of system-level policy-amenable variables with reading literacy performance. *Educational Research and Evaluation*, *21*(1), 15-39.

Schweisfurth, M., & Elliott, J. (2019). When 'best practice' meets the pedagogical nexus: recontextualisation, reframing and resilience. *Comparative Education*, *55*(1), 1-8.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, *18*(2-3), 277–310.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher Professional Learning and Development: Best Evidence Synthesis Iteration [BES]*. New Zealand: New Zealand Ministry of Education.

TIMSS (2015). Available online at: https://www.iea.nl/studies/iea/timss/2015 (Accessed 26th July 2019).

Turner, H., Boruch, R., Petrosino, A., Lavenberg, J., De Moya, D., & Rothstein, H. (2003). Populating an international web-based randomized trials register in the social, behavioral, criminological, and education sciences. *The ANNALS of the American Academy of Political and Social Science*, *589*(1), 203-223.

Tymms, P., & Merrell, C. (2009). Standards and Quality in English Primary Schools Over Time. *The Cambridge Primary Review Research Surveys. R. Alexander, C. Doddington, J. Gray and L. Hargreaves.*

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, *30*(4), 477-494.

Tymms, P. (2011). The Impact of Large-Scale Reform in England. *Zeitschrift für Erziehungswissenschaft*, *13*(1), 105-116.

Tymms, P., Merrell, C., & Wildy, H. (2015). The progress of pupils in their first school year across classes and educational systems. *British Educational Research Journal*, *41*(3), 365-380.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2009-455. *National Center for Education Statistics*.

Whetton, C. (2009). A brief history of a testing time: National curriculum assessment in England 1989–2008. *Educational Research*, *51*(2), 137-159.

Weick, K. E. (1976). Educational organisations as loosely coupled systems. *Administrative Science Quarterly*, *21*(1), 1–19.

Wiseman, A.W., & Chase-Mayoral, A. (2014). Shifting the discourse on neo-institutional theory in comparative and international education, in: A.W. Wiseman & E. Anderson (Eds) *Annual review of comparative and international education 2013*. Bingley: Emerald Group Publishing Ltd.

Standards in Education: Reforms, Stagnation and the Need to Rethink

David Bolden[a*] and Peter Tymms[a]
[a]School of Education, Durham University, Durham, UK, DH1 1 TA

Fig 1. Key Stage 2 test results for mathematics and English (achieving Level 4 or above) over time.

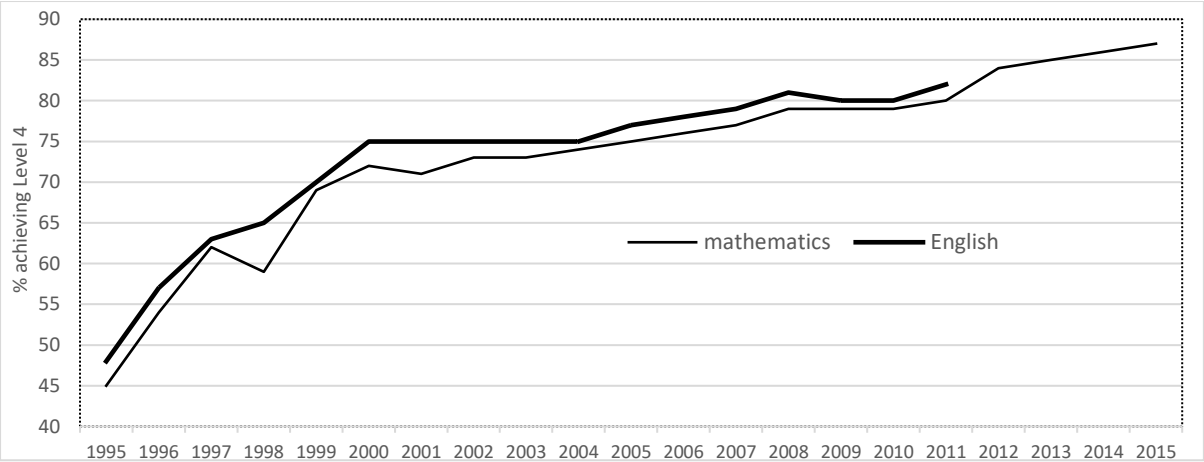
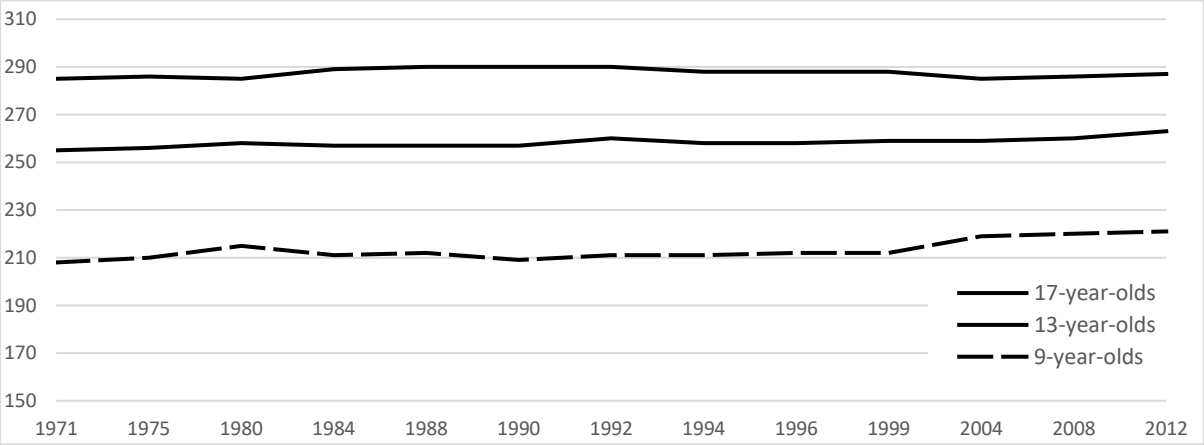


Fig. 2. Trend in average reading scores by age from 1971 to 2012[1] (National Assessment of Education Performance).



[1] Some changes to testing procedures occurred in 2004 in both the reading and mathematics to accommodate students with disabilities and EAL.

Fig. 3. Long-term trend in average mathematics scores by age from 1973 to 2012 (National Assessment of Education Performance).

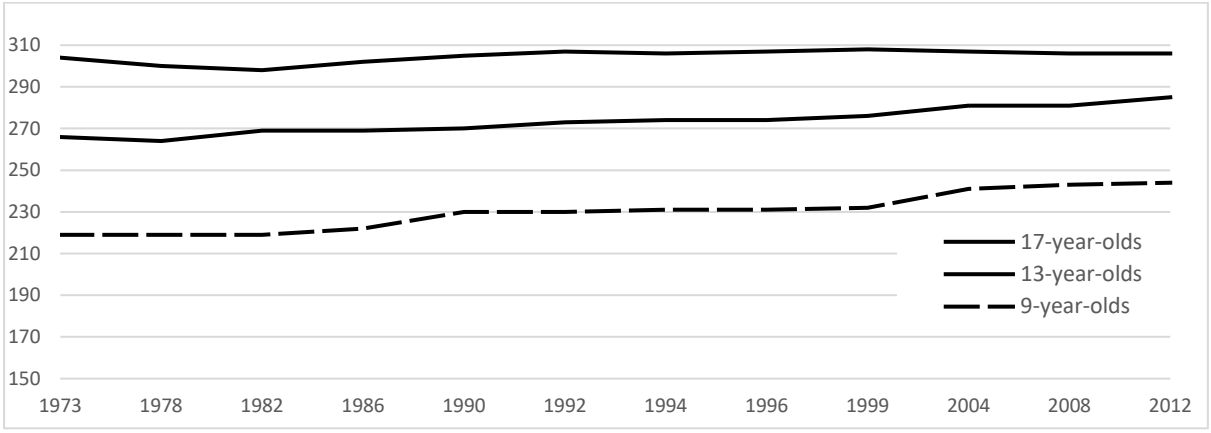


Fig 4. Trends in average mathematics scores for those countries involved in all TIMSS cycles by grade 4 (11 countries plus 2 benchmarking) and grade 8 (10 countries plus 2 benchmarking) (* denotes no mathematics data for Grade 4 in 1999).

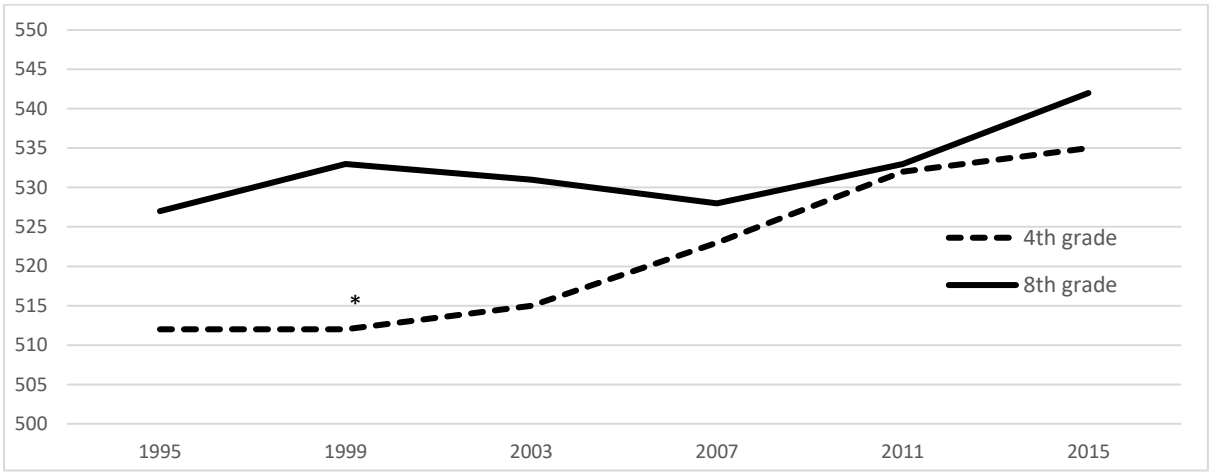


Fig 5. Trends in science scores for those countries involved in all TIMSS cycles by grade 4 (11 countries plus 2 benchmarking) and grade 8 (10 countries plus 2 benchmarking) (* denotes no science data for Grade 4 in 1999).

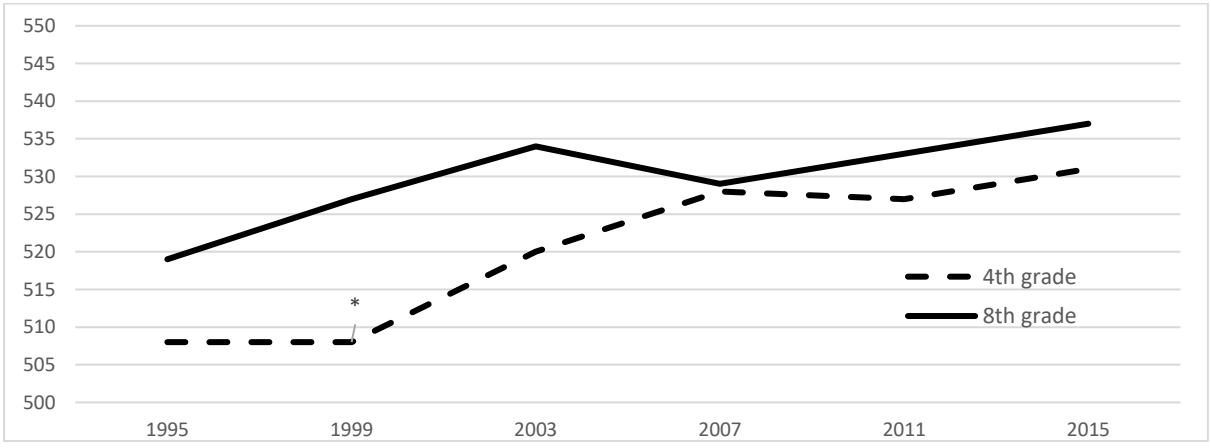


Fig 6. Trends in average PIRLS scores for countries participating in all cycles.