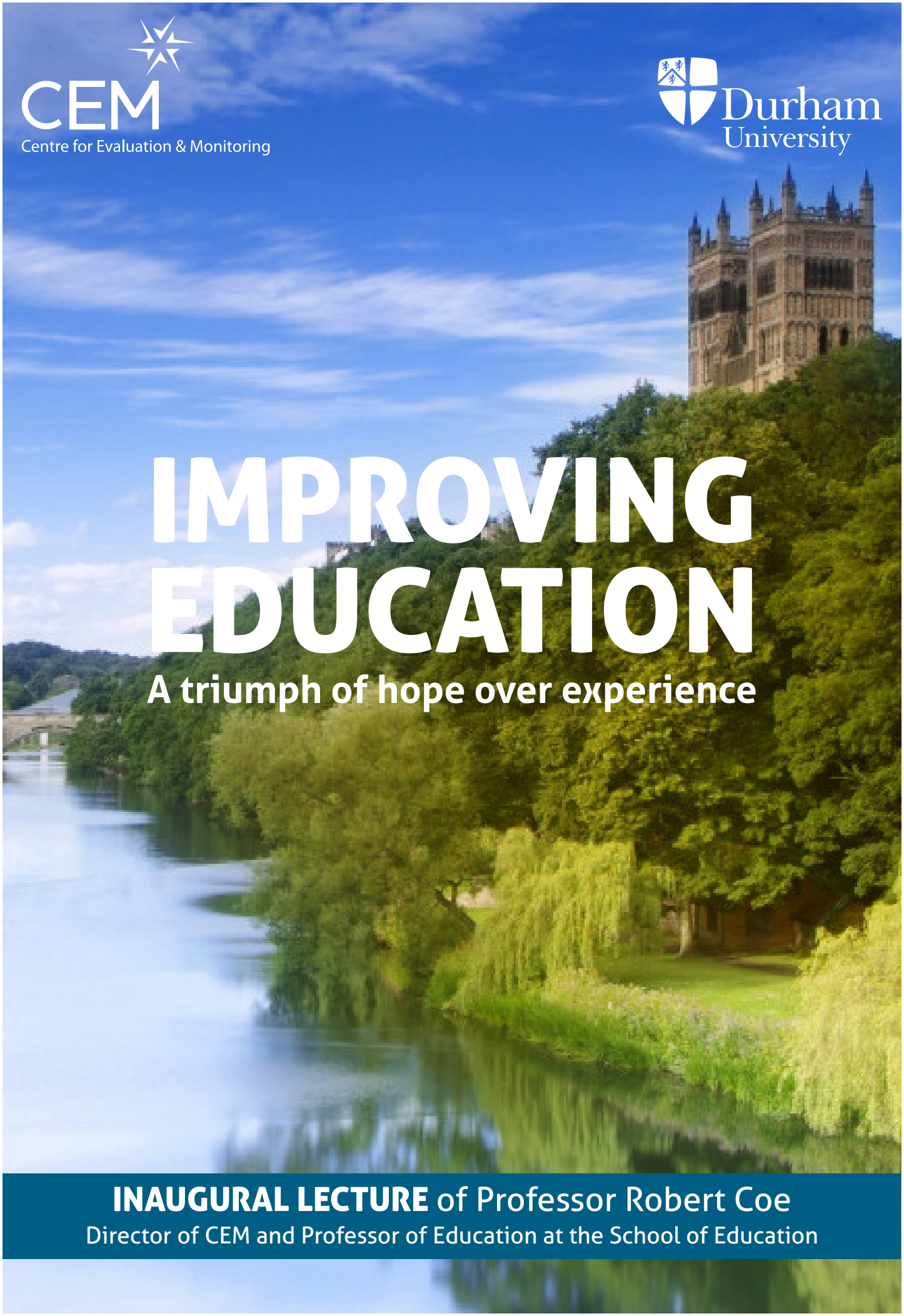

CEM
Centre for Evaluation & Monitoring
Durham University
IMPROVING EDUCATION
A triumph of hope over experience
INAUGURAL LECTURE of Professor Robert Coe
Director of CEM and Professor of Education at the School of Education

Improving Education: A triumph of hope over experience

Inaugural Lecture of Professor Robert Coe,
Durham University,
18 June 2013

Despite the apparently plausible and widespread belief to the contrary, the evidence that levels of attainment in schools in England have systematically improved over the last 30 years is unconvincing. Much of what is claimed as school improvement is illusory, and many of the most commonly advocated strategies for improvement are not robustly proven to work. Even the claims of school effectiveness research – that we can identify good schools and teachers, and the practices that make them good – seem not to stand up to critical scrutiny. Recent growth of interest in evidence-based practice and policy appears to offer a way forward; but the evidence from past attempts to implement evidence-based approaches is rather disappointing. Overall, an honest and critical appraisal of our experience of trying to improve education is that, despite the best intentions and huge investment, we have failed – so far – to achieve it.

Nevertheless, we have reason to be hopeful, provided we are willing to learn from our experience. Specifically, I will argue that we need to do four things: to be clear what kinds of learning we value; to evaluate, and measure properly, teaching quality; to invest in high-quality professional development; and to evaluate robustly the impact of changes we make.

Introduction

I am a school teacher who became interested in educational research, and ended up doing that instead. Most of my research has looked at questions that are directly relevant to practice or policy: I want to make education better for children and young people. I know there is nothing special about that; the hard question is: How?

On one level, my analysis is quite bleak: standards have not risen; teaching has not improved; research that has tried to support improvement has generally not succeeded; even identifying which schools and teachers are good is more difficult than we thought. That is our experience, so far. Recognising this is important, not because I enjoy puncturing inflated assertions of success and what seem to me to be complacent and uncritical claims of 'fools

gold' improvement (although, I confess, I do), but because I think it is time we stopped repeating the same mistakes.

I am optimistic about our capacity to learn from what has not worked, not to keep on repeating the same mistakes, and to use what knowledge we have about what seems most likely to make a difference. Most of all, I believe in the power of evaluation to tell us what is working and of feedback loops to allow that evaluation to influence practice. Despite our experience, there are strong grounds for hope that we can do better.

Experience

First I will talk about the past, and ask whether what we have done so far has worked.

Have educational standards really risen?

This is inevitably controversial; my answer will surely offend many people who have invested energy and resources into trying to raise standards. Those who have invested huge energy and resources may be hugely offended. So why do I feel the need to say things that I know will offend them?

I want to be clear that I do not mean to imply any criticism of teachers or anyone else working in education. I was a teacher myself and I know how hard teachers work, how committed they are to doing the best for, and getting the best from, their students, no matter what challenges they face.

However, if it is true that despite the huge efforts we have made to improve education not much has changed, there are important lessons for us to learn. One would be that effort and good intentions are not enough; we have to work smarter, not just harder. Another would be that we must look carefully at the strategies we have been using to improve, and replace them with some different ones. A third lesson is that a more critical and realistic approach to evaluation may be required. An uncritical belief that things are improving may be comforting, but is ultimately self-deceiving and unproductive.

In short, I find it hard to see how we can make real improvement until we accept the unpalatable truth that we have so far failed to achieve it.

Unfortunately, a clear and definitive answer to the question of whether standards have risen is not possible. The best I think we can say is that overall there probably has not been much change. However, we are limited by the fact that in England there has been no systematic, rigorous collection of high-quality data on attainment that could answer the question about systemic changes in standards. And we might well disagree about what we mean by 'standards' or how they should be measured. The evidence we have is patchy and inadequate, but it is the best we have. There are three main types of evidence: international surveys, independent studies and national examinations.

Evidence from international surveys

International surveys of attainment, such as PISA, PIRLS and TIMSS, have become increasingly prominent as indicators of our standing in the international league table of performance. These studies are not designed to evaluate change over time and their interpretation and use is inevitably problematic (see e.g. Brown, 1998; Jerrim, 2011). Nevertheless, they provide a source of evidence, based on a high-quality assessment development process, robust attempts to equate scores across testing occasions and a rigorous national sampling process.

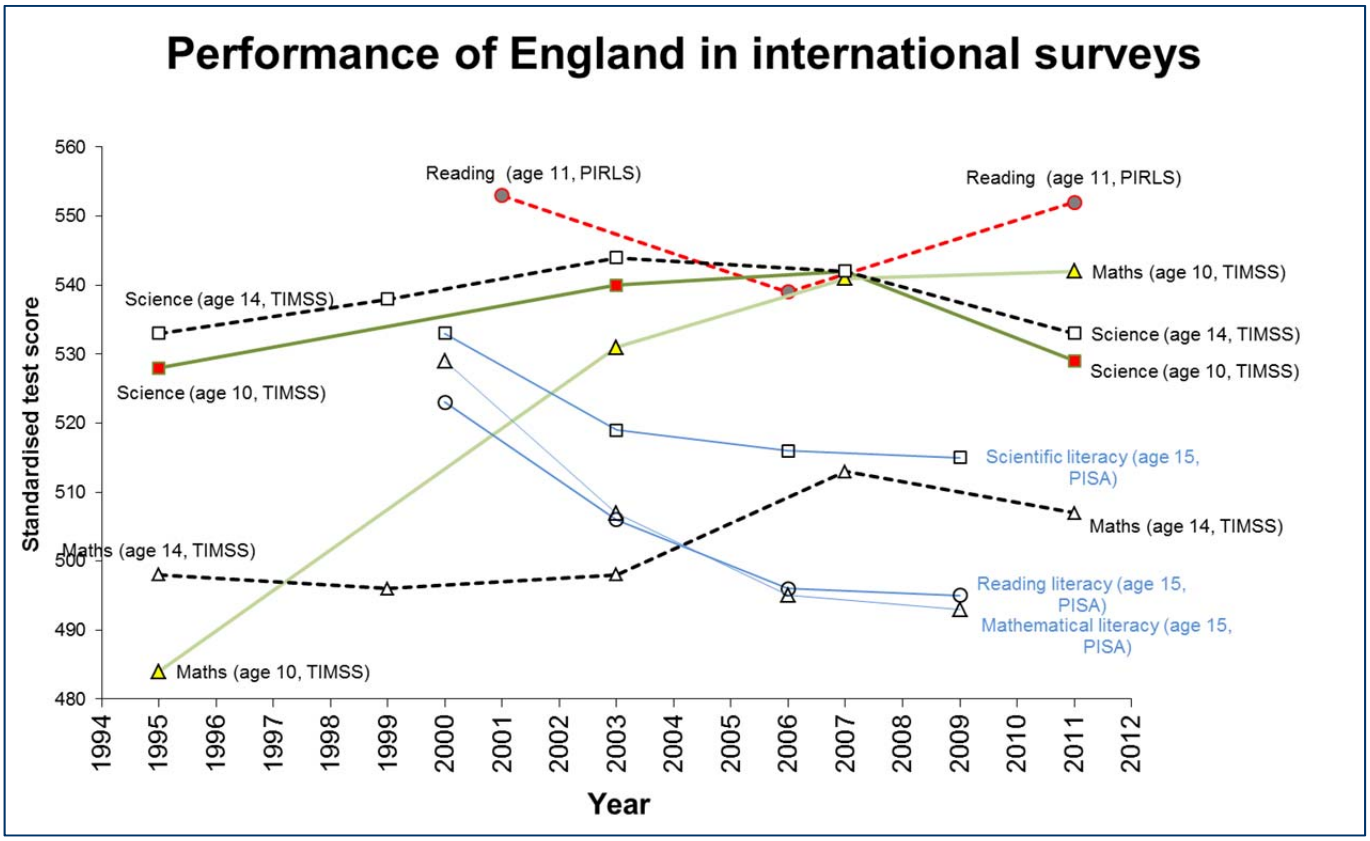
Figure 1 shows the scaled average performance for England in PISA, PIRLS and TIMSS between 1995 and 2011. It is a mixed picture. In some curriculum areas, for some ages, performance has improved (e.g. maths age 10 in

TIMSS). For others, it has declined (e.g. PISA, age 15, all subjects). There isn't really a coherent story here. There might be a case for saying that gains at age 10 are of little value if they are lost by age 14 or 15, hence focusing on the results for older children. For them, TIMSS suggests little change, PISA seems to show a fall.

Just to put these results in context, a change of 25 points in PISA has been estimated to be worth £4 trillion to England's GDP (Hanushek and Woessmann, 2010) and would typically be the difference between an average performing country (such as England) and one ranked 5-10 internationally (such as Canada or New Zealand) (OECD, 2010). The gap between the average and the top performing country is typically 50-60 points.[1]

However, when official sources (a DfE press release and comments from Sir Michael Willshaw, Chief Inspector) cited the PISA decline as evidence of falling standards, they were censured by the UK Statistics Agency for making comparisons that were 'statistically problematic' (Stewart, 2012), partly because the response rate in England in 2000 and 2003 had been low. An analysis by Jerrim (2011) tried to estimate the effects of higher non-response for England on PISA in 2000 and 2003, as well as for the change in test month and – an important but little known change – the fact that in 2006 & 2009 it was only Y11 15-year olds who were included (in 2000 & 2003 all 15-year olds, whether in Y10 or Y11, were sampled). Overall, these corrections suggested there had probably not been much change between 2000 and 2009.

Figure 1: Performance of England in international surveys



Hence, overall, it may be that the pattern of results from the different international surveys is actually fairly consistent: not much change between 1995 and 2011.

[1] Shanghai-China is an outlier, with, for example, an average score of 600 in mathematics in PISA 2009 (ie about 100 above the average)

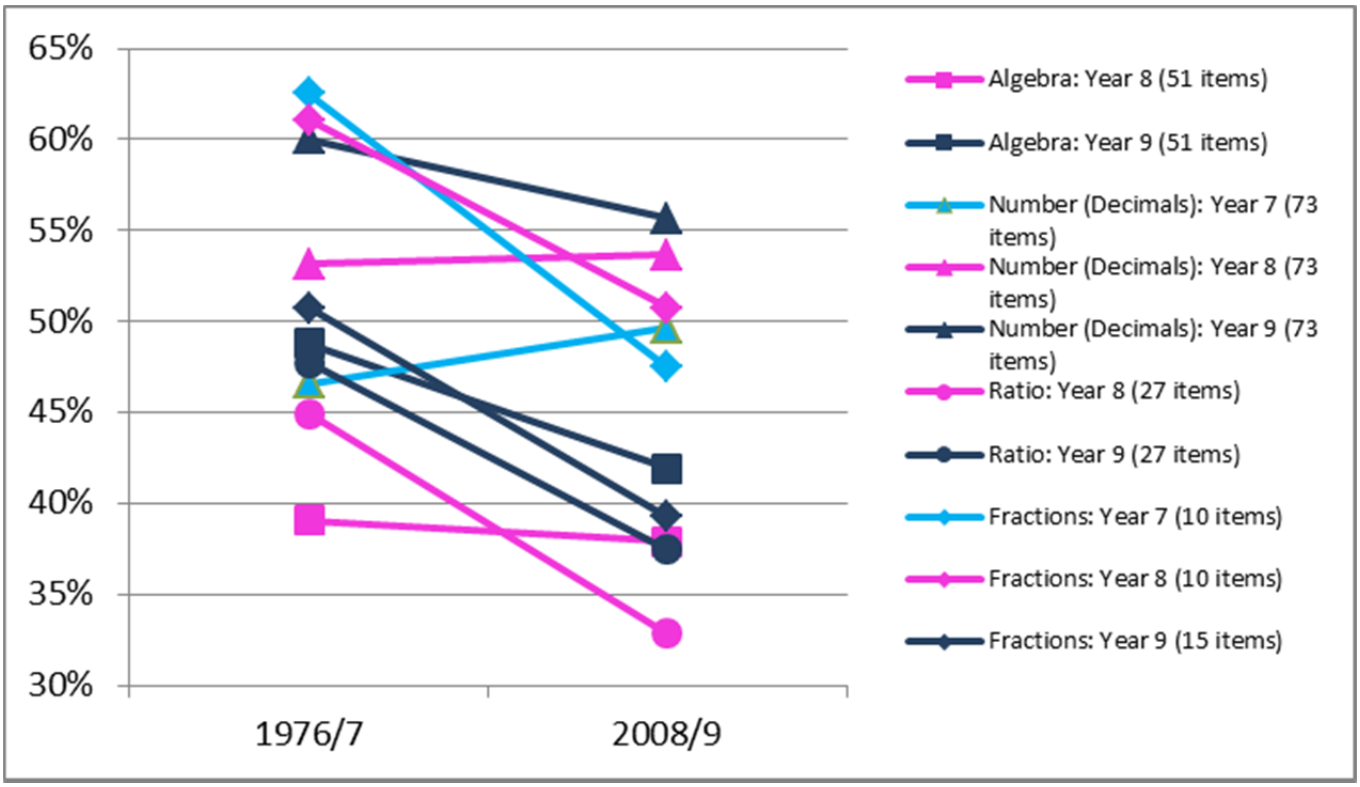
Evidence from independent studies

Some independent studies have presented data that may be interpreted as evidence of national changes. The following brief review is far from comprehensive or systematically collected, so it may well be that the addition of other studies could change the balance of results.[2]

Tymms and Merrell (2007) reviewed assessment evidence from children at the end of primary school in England. They examined a range of assessment data in mathematics between 1978 and 2004 and in reading between 1948 and 2004. Overall, they concluded that standards in both subjects 'have remained fairly constant' over most of this period, but more recently (i.e. since 1995), 'Reading has risen very slightly and mathematics has risen moderately' (p5).

Hodgen et al (2009, 2010, 2012) conducted a survey of mathematical performance in 2008-9 with a nationally representative sample, and compared the results with performance on the same tests in 1976-7. The results are summarised in Figure 2. Although there were rises in Number for pupils in in Y7 and Y8, for every other aspect of mathematics, and for Number in Y9, the overall pattern was of decline.

Figure 2: Changes in performance on the ICCAMS / CSMS mathematics assessments between 1976/7 and 2008/9



Coe and Tymms (2008) reported changes in the performance of 14-year-old students (at the start of Y10) in the Yellis sample on a test of mathematics and vocabulary between 1998 and 2007. Overall, they found a 'small but steady rise', equivalent to 0.3 of a standard deviation, over that period.

By contrast, Shayer et al (2007) found that performance in a test of basic scientific concepts seemed to have fallen sharply between 1976 and 2003.

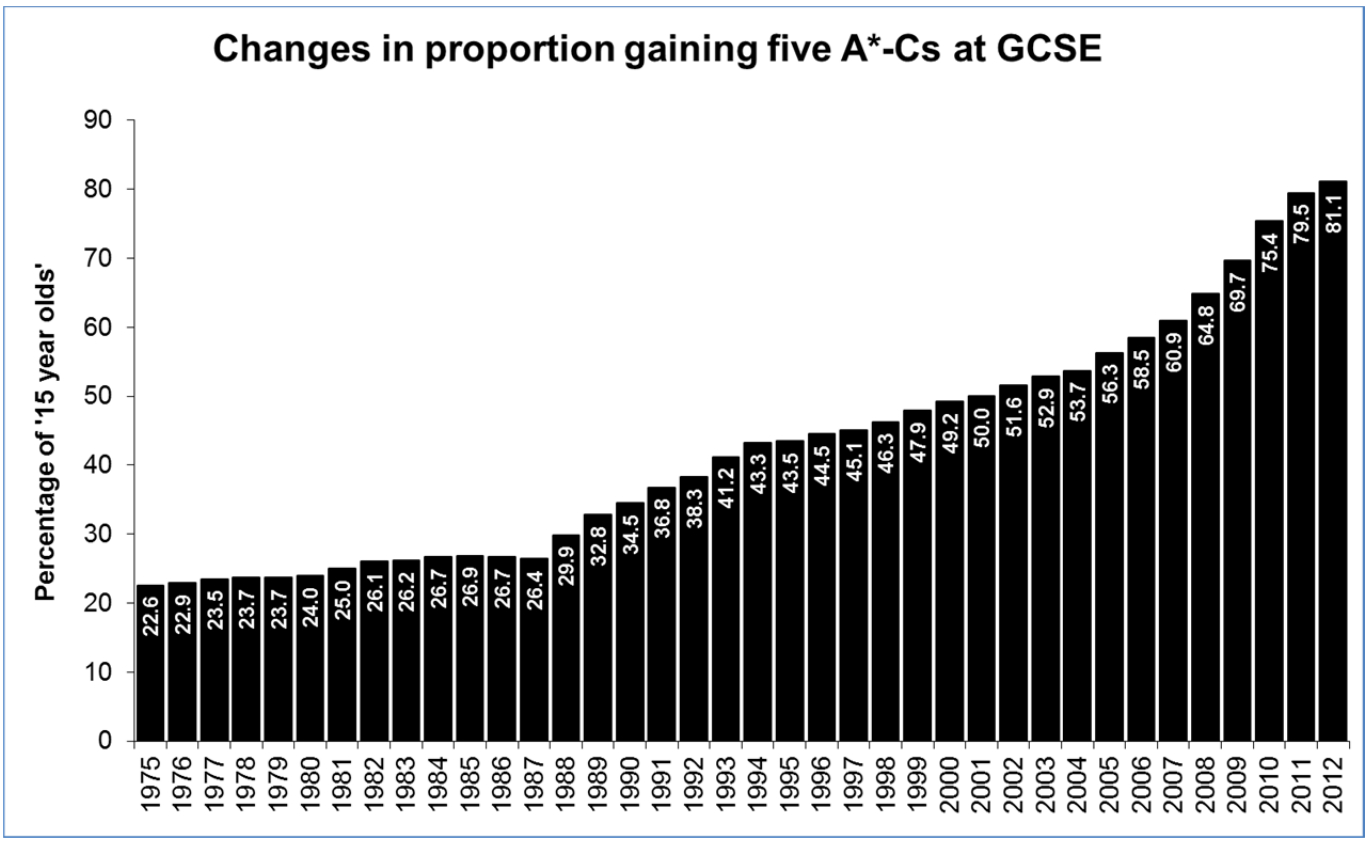
[2] If anyone knows of any further studies, I would be delighted to hear of them.

Evidence from national examinations

The most obvious support for claims that educational standards in England have risen draws on the dramatic rises in national performance at Key Stage 4 (GCSE, age 16) and Key Stage 2 (Y6 SATs, age 11). For GCSE, a picture of the change between 1975 and 2012 is shown in Figure 3. GCSEs are taken by almost all school students in England; their grading is subject to intense scrutiny and regulation by publicly accountable bodies who are charged with maintaining the standard from one year to the next. Any changes in grades achieved must therefore be taken seriously as indicative of changes in educational standards.

When GCSE was introduced in 1987, 26.4% of the cohort achieved five grade Cs or better. By 2012 the proportion had risen to 81.1%. This increase is equivalent to a standardised effect size of 1.63, [3] or 163 points on the PISA scale, which is twice the full vertical axis shown on Figure 1. If we limit the period to 1995 – 2011 (as in Figure 1) the rise (from 44% to 80% 5A*-C) is equivalent to 99 points on the PISA scale. In Figure 4 this rise is superimposed on the data from Figure 1.

Figure 3: Changes in proportion gaining five A-Cs at GCSE*



Sources: SCAA (1996); www.gov.uk (various URLs that are never the same twice)

It is clear from Figure 4 that the two sets of data tell stories that are not remotely compatible. Even half the improvement that is entailed in the rise in GCSE performance would have lifted England from being an average performing OECD country to being comfortably the best in the world. To have doubled that rise in 16 years is just not believable.

[3] Conversion of percentage achieving a threshold to a standardized effect size is based on a probit function.

Figure 4: Changes in performance of England in international surveys, compared with change in GCSE performance

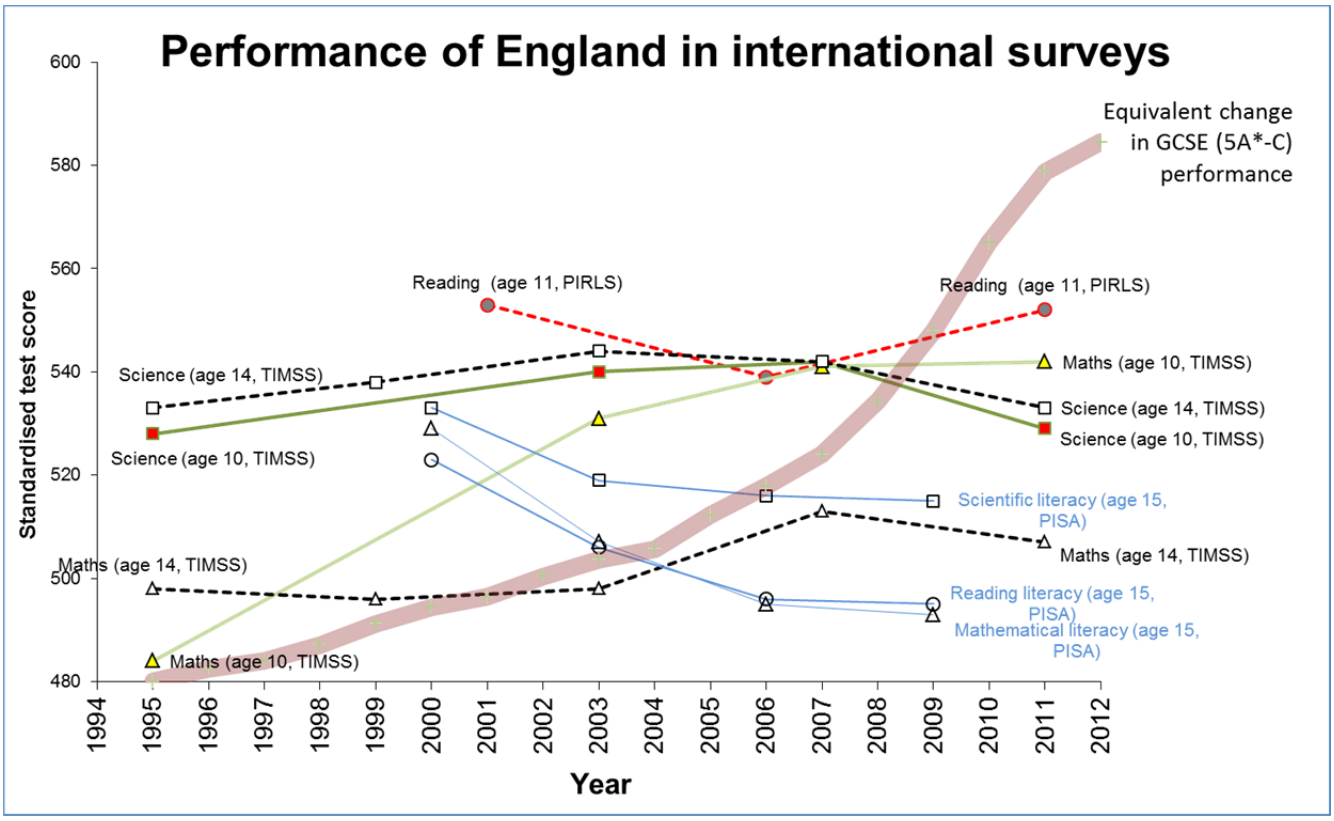
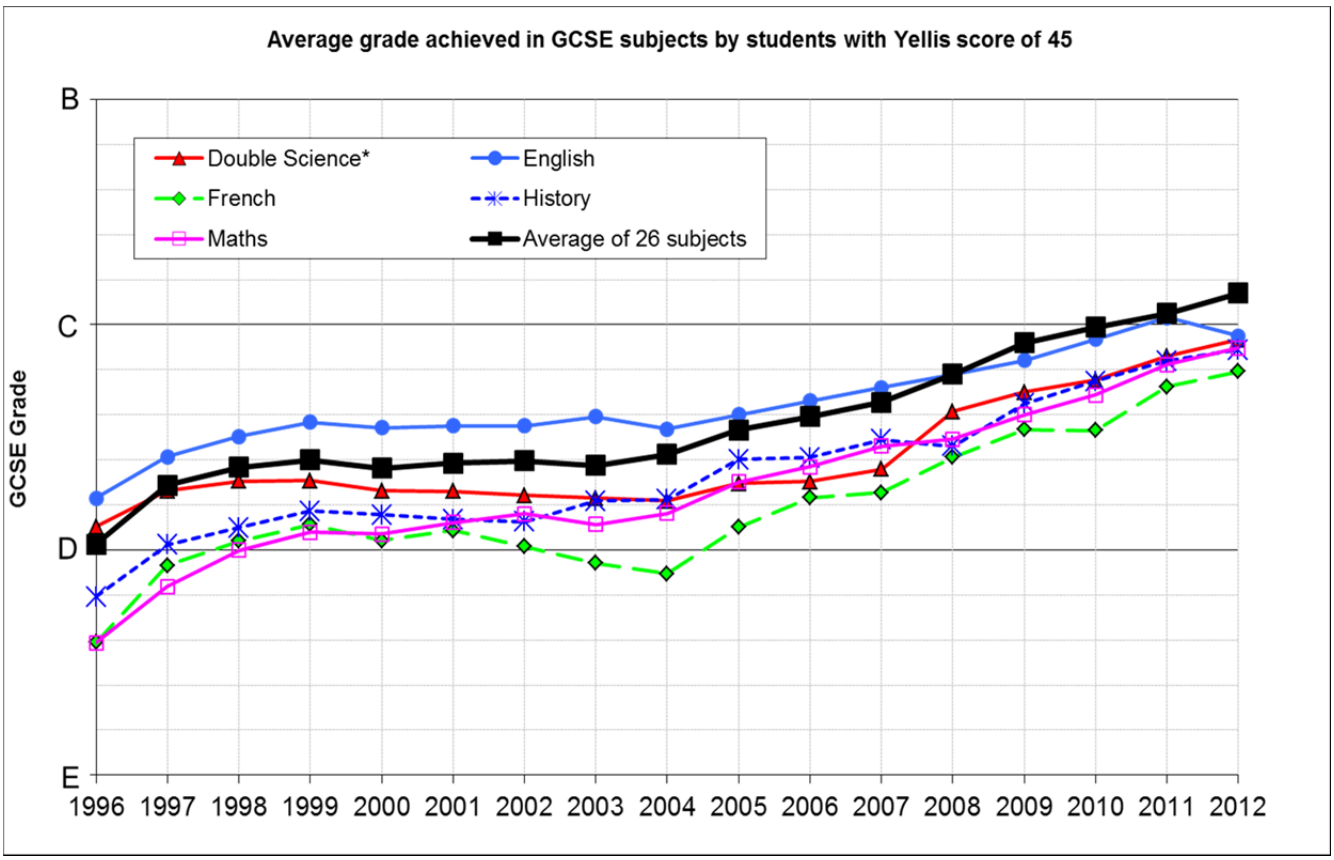


Figure 5: Changes in GCSE grades achieved by candidates with the same maths and vocabulary scores each year



The question, therefore, is not whether there has been grade inflation, but how much. Even if some of the rise is a result of slipping grade standards, it is still possible that some of it represents genuine improvement. The extent to which rising GCSE grades can be interpreted as evidence of rising standards seems very problematic, however.

Further evidence of grade inflation comes from my (Coe, 2007) analysis of the GCSE grades achieved by candidates of the same ability each year. Ability here is measured by the same Yellis test of maths and vocabulary cited above. An update of this analysis, to include GCSE results up to 2012, is presented in Figure 5.

It is not straightforward to interpret the rise in grades in Figure 5 as grade inflation; a range of alternative explanations are put forward by Coe (2007). However, it does suggest that whatever improved grades may indicate, they do not correspond with improved performance in a fixed test of maths and vocabulary.

School improvement: Isn't it time there was some?

If standards have not really risen then it surely follows that there cannot have been systemic school improvement. Of course, some schools will have improved, but perhaps just as many will have declined. Once again, the assertion that, overall, schools have not improved seems to be at odds with the prevalence of claims about improving schools from a wide range of sources.

Figure 6: Mistaking School Improvement

Mistaking School Improvement:
How to make it look as if your improvement project has worked

1. Wait for a bad year or choose underperforming schools to start with. Most things self-correct or revert to expectations (you can claim the credit for this).
2. Take on any initiative, and ask everyone who put effort into it whether they feel it worked. No-one wants to feel their effort was wasted.
3. Define 'improvement' in terms of perceptions and ratings of teachers. DO NOT conduct any proper assessments – they may disappoint.
4. Only study schools or teachers that recognise a problem and are prepared to take on an initiative. They'll probably improve whatever you do.
5. Conduct some kind of evaluation, but don't let the design be too good – poor quality evaluations are much more likely to show positive results.
6. If any improvement occurs in any aspect of performance, focus attention on that rather than on any areas or schools that have not improved or got worse (don't mention them!).
7. Put some effort into marketing and presentation of the school. Once you start to recruit better students, things will improve.

In another paper (Coe, 2009), I highlighted some of the weaknesses of claims in the research literature about school improvement. I listed some of the main inadequacies of evaluation design and interpretation in the form

of guidance a mistaken school improvement consultant might offer to a school or school system that wants to be able to claim 'improvement'. Figure 6 is based on that list.

Until it becomes routine for school improvement initiatives to be evaluated using designs that rule out at least these explanations, the state of our knowledge about how to improve schools will remain limited.

Can we identify effective schools and teachers?

In the same paper (Coe, 2009) and elsewhere (Coe and Fitz-Gibbon, 1998; Dumay, Coe & Anumendem, 2013) I have argued that the interpretation of residual gains in value-added models of school effectiveness as the impact of the school or teacher on learning is a lot more problematic than generally recognised. Others have also argued this (e.g. Hill, 1998; Ouston, 1999; Scheerens et al., 2001; Luyten et al., 2005; Gorard, 2010).

There are two key questions that research in this area needs to address before we should accept its claims that its methods really do identify effective schools or teachers:

1. Can we interpret value-added (or other ways of calculating 'effectiveness') as a causal effect of the school or teacher?

2. If so, are the characteristics associated with that effectiveness alterable in ways that lead (causally) to improvement?

Is 'value-added' the same as 'effectiveness'?

It is a commonplace of research methods training that correlation is not causation; causal claims need careful causal arguments and evidence to support them (Shadish, Cook and Campbell, 2002). Another basic principle of methodology is that measures derived from an instrument or process should not be interpreted as having a particular meaning until a proper validity argument has been made (Kane, 2006). It therefore seems to be an oddly unscrutinised anomaly that research in school effectiveness has routinely interpreted the gains in students' test scores as the 'effect' of the school or teacher.

Since 1983, the Centre for Evaluation and Monitoring (CEM) at Durham University[4] has been providing assessment, analysis and monitoring systems for schools to help educators evaluate their own performance. Thirty years ago many of the underpinning ideas were revolutionary: value-added analysis; use of regular, robust assessment; tracking individual pupils' progress; evaluating the performance of schools, departments and individual teachers. Now, in part due to CEM's influence, these ideas are mainstream.

From working with schools and teachers interpreting value-added scores of their students for 30 years, we know that value-added is not always the same as effectiveness. For example, one of the simplest ways for a teacher to achieve high value-added scores is to follow a teacher whose value-added is low. Another is to teach top sets, selected for their likelihood of attaining high grades.[5] Other reasons why students choose or are placed in different teaching groups can make more difference to their value-added than who the teacher is. Our general advice is summed up in the maxim: 'Value-added data may not answer your questions, but it can help you to ask better ones'.

[4] CEM was founded at Newcastle University and moved to Durham in 1996. It was originally called the *Curriculum, Evaluation and Management Centre* and the hard 'C' of Curriculum established its pronunciation as 'KEM'. In 2007 the lack of any real connection with either curriculum or management led us to change the name in a re-brand to the more appropriate *Centre for Evaluation and Monitoring*, keeping the same acronym (CEM) and pronunciation. See www.cem.org

[5] See http://www.cem.org/attachments/081_Value-Added%20and%20teaching%20sets%20Nov%202002.pdf

Characteristics of 'effectiveness' as levers for change?

Having made the assumption that learning gains can be interpreted as 'effectiveness', school effectiveness research then often goes on to list the characteristics of schools or teachers that are associated with them. Characteristics like 'strong leadership', 'high expectations', 'positive climate' and a 'focus on teaching and learning' are typically found. Never mind that these characteristics are often only vaguely defined (Coe and Fitz-Gibbon, 1998) and that these correlations are generally pretty low (Scheerens, 2000, 2012): does knowing them actually help us?

If we knew that there were clear strategies for helping schools or teachers that did not have those characteristics to adopt them that would be a start. Then if we could show that implementing these strategies did in fact increase the prevalence of the target characteristics we might have a hypothesis. But only if after that a series of robust evaluation studies showed that this did in fact lead to improvements in student attainment could we really claim that this knowledge was helpful.

I would be delighted to be contradicted on this, but I don't think this is the case for any of the much-touted characteristics of effectiveness.

Is 'evidence-based' practice and policy the answer?

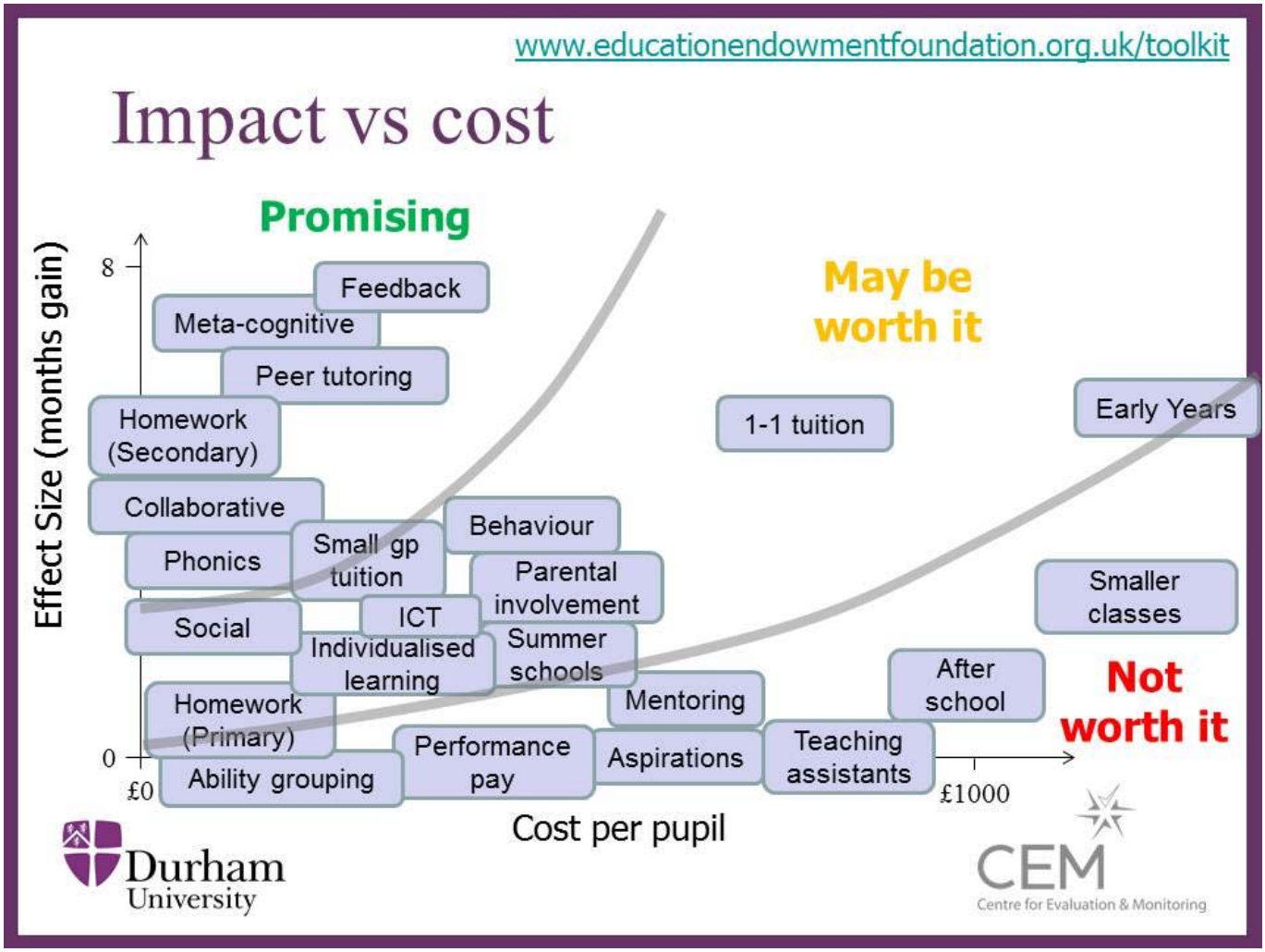
A small group of UK educational researchers have been trying to promote 'evidence-based' education for some time; many of these are associated with Durham (e.g. Fitz-Gibbon, 1996; Davies, 1999; Coe, Fitz-Gibbon and Tymms, 2000; Torgerson & Torgerson, 2001; Gorard and Cook, 2007; Tymms, Merrell and Coe, 2008; see also http://www.cem.org/evidence-based-education/previous-events).

More recently, an accessible review of the relative cost and benefit of different approaches, written for teachers, commissioned by the Sutton Trust and now supported by the Educational Endowment Foundation, has had significant influence on policy and practice (Higgins et al, 2013). A graphical summary of the results reported in this *Toolkit* is shown in Figure 7.

Despite the fact that the *Toolkit* merely summarises research findings, most of which have been known for a long time, some of the results seem to have been surprising and controversial. For example, some things that are popular or widely thought to be effective seem, on the basis of this evidence, to be probably not worth doing. Strategies such as setting learners in ability groups, providing after-school clubs, employing teaching assistants, reducing class sizes, introducing performance pay for teachers and focussing on raising aspirations are all either ineffective or not effective enough to justify their costs, in terms of their impact on learning outcomes.

More positively, some other strategies look promising: giving effective feedback; encouraging meta-cognitive and self-regulation strategies; using peer tutoring/peer-assisted learning; setting homework (for secondary age pupils); promoting collaborative learning; and the use of phonics in learning to read.

Figure 7: Summary of cost and impact estimates from the Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit



On the face of it, the advice that should be given to teachers is clear and simple: choose the strategies that have maximum impact for their cost; avoid those that have no impact, or small impact and high cost. In short, choose from the top left. But, in the quotable words of H.L. Mencken, 'For every complex problem there is an answer that is clear, simple, and wrong'.

Our suspicions that it is not as simple as this should be alerted by the fact that we have already been doing some of these things for a long time, but, as argued above, they have generally not worked. An example is Assessment for Learning, which incorporates many of the key principles of the two most promising strategies: feedback and meta-cognitive regulation. These strategies have been around for many years, but became the focus of national policy, widely endorsed by teachers and supported by extensive government training, following the publication of Black and Wiliam's (1998) *Inside the Black Box* (see Black and Wiliam, 2003, for an account). It is now a rare thing, in my experience, to meet any teacher in any school in England who would not claim to be doing Assessment for Learning. And yet, the evidence presented above suggests that during the fifteen years of this intensive intervention to promote AfL, despite its near universal adoption and strong research evidence of substantial impact on attainment, there has been no (or at best limited) effect on learning outcomes nationally.

How can we explain this? Why is this clear and simple advice wrong? I think here are two main reasons.

Research evidence is problematic

The first reason is that the evidence itself is sometimes not as secure or generalisable as we might like. For example, for some of the interventions reviewed, the evidence is limited in quality, quantity or relevance. This is acknowledged in the *Toolkit*, but it is inevitably a complex judgement in each case how clear, convincing and applicable the findings are.

Related to this is the need to acknowledge that the results of reviewing research studies of the impact of interventions may not correspond with the likely impacts of making those changes at scale in real contexts. We know, for example, that positive results are more likely to be published and hence to be included in reviews (Ioannidis, 2005). We also know that small studies and studies where the evaluator is also the developer or deliverer of the intervention tend to report larger effects than large-scale evaluations where there is separation of roles (Slavin and Smith, 2009); the latter are probably more likely to represent the impact if the intervention is implemented in real schools.

Another related issue is that effects often depend on a combination of contextual and 'support factors' (Cartwright and Hardie, 2012) that are not always understood. Sometimes things that are proven to work turn out not to. Evaluations tell us what did work there; they do not always guarantee that if we try to do the same it will work here too.

Implementation is problematic

The second main problem with just advocating the high-impact strategies presents perhaps even more of a challenge. Many of the most effective strategies are complex, open to interpretation and hard to implement. We may think we are doing it, but are we doing it right? In most cases the approach is not supported by a well-defined, feasible, large-scale intervention strategy. In other words, we do not know how to get large groups of teachers and schools to implement these interventions in ways that are faithful, effective and sustainable.

Hope

My argument so far has not been encouraging: things are not getting better, despite the best intentions and huge effort. Improving school systems is clearly very hard. Anyone who is convinced by the argument up to this point may be forgiven for thinking that the hope that we can do better in the future does not seem to be grounded in evidence.

This part of my argument is not really evidence-based. In all honesty I do not know if we can do better; but I'm just not ready to give up yet. If we keep trying things that seem most plausibly likely to improve the outcomes we value, and – crucially – keep evaluating our progress, then eventually we must surely learn to make progress.

So what should we do (that hasn't failed yet)?

Our strategy should therefore be to make the best choices we can from the best evidence available, to try it out, with an open mind, and see if it works. If it does, we can keep doing it; if not, we will learn from that experience and try something else.

I think there are four strategies that are worth trying.

1. Think hard about learning

The success of the *Toolkit* in generating interest in evidence-based strategies has led to many invitations for me to talk about it to groups of teachers. When I get to the bit about the strategies that are popular with many teachers but the evidence suggests are not effective enough to justify their costs (e.g. reducing class size, employing teaching assistants, ability grouping) the discussion is often quite challenging. When our intuition is in conflict with research evidence we have to make a choice. Sometimes the research turns out to be wrong and we are right to hold on to our intuitions; but the history of science is full of examples of conflicts between intuition and evidence, and mostly the evidence wins. As the evidence comes to be accepted, our intuitions and understandings change and become more sophisticated in order to accommodate it. This process of challenging and reconstructing our understandings or schemas in the light of new evidence or experience is also part of many theories of learning.

Part of the challenge for me is to understand why these 'ineffective' approaches are so strongly believed to benefit learning by many of those whose experience is rooted in the classroom. The best answer I have come up with is that those strategies do indeed have benefits, but for outcomes that we mistake for learning rather than for real learning. In other words, the teachers who believe strongly and unshakably that reducing class size from 30 to 15, or adding a teaching assistant into a mainstream classroom, or setting learners into ability groups will make a big difference to learning are often justifying this belief by drawing on an understanding of 'learning' that is at odds with the one being measured in research studies.

For example, in discussing why they believe smaller classes are much better, a teacher will often say, 'You can give pupils more individual attention.' My question is then, 'Does more individual teacher attention mean more learning? What makes you think that?' In fact, much of the learning that happens in classrooms can be unexpectedly unrelated to what teachers intend, assume or do (Nuthall, 2005, 2007). But somehow the idea that 'I have taught it' becomes a proxy for 'they have learned it', without a need for any independent check on what (if anything) has actually been learned.

Figure 8: Poor Proxies for Learning

Poor Proxies for Learning

(Easily observed, but not really about learning)

1. Students are busy: lots of work is done (especially written work)
2. Students are engaged, interested, motivated
3. Students are getting attention: feedback, explanations
4. Classroom is ordered, calm, under control
5. Curriculum has been 'covered' (ie presented to students in some form)
6. (At least some) students have supplied correct answers (whether or not they really understood them or could reproduce them independently)

Doing justice to this kind of argument needs a lot more time and space than is available here. In discussing, thinking and reading about these issues, and in observing classrooms, I have come to believe that it may be common for teaching to be quite unrelated to learning; in many classrooms both things are happening, but each

is more or less independent of the other. This can only happen if teaching is not really focussed on learning, and the reason that is possible is that teachers readily accept poor proxies for learning, rather than seeking direct and valid evidence of true learning (which is much harder!). Some of these poor proxies are listed in Figure 8.

If it is true that teaching is sometimes not focussed on learning, how can we make them better aligned? One answer is that it may help to clarify exactly what we think learning is and how it happens, so that we can move beyond the proxies. I have come up with a simple formulation:

Learning happens when people have to think hard

Obviously, this is over-simplistic, vague and not original. But if it helps teachers to ask questions like, 'Where in this lesson will students have to think hard?' it may be useful.

Some research evidence, along with more anecdotal experience, also suggests that students themselves may not necessarily have real learning at the top of their agenda either. For example, Nuthall (2005) reports a study in which most students "were thinking about how to get finished quickly or how to get the answer with the least possible effort". If given the choice between copying out a set of correct answers, with no effort, but no understanding of how to get them, and having to think hard to derive their own answers, check them, correct them and try to develop their own understanding of the underlying logic behind them, how many students would freely choose the latter? And yet, by choosing the former, they are effectively saying, 'I am not interested in learning.'

This leads me to my first recommendation: **Get teaching really focussed on learning**

The hard question here is: How? Which brings me to my second strategy.

2. Invest in effective professional development

There seems to be a lot of interest in teachers' continuing professional development (CPD) at the moment, and a lot of this argument has been well made by others (eg Teacher Development Trust, 2013).

We know that teaching quality is one of the main determinants of learning (Hattie, 2003; Rivkin et al, 2005; Kane et al, 2013). We also know that measures of teachers' performance typically improve during the first few years of their careers, and then plateau after 3-5 years (Rockoff, 2004; Kukla-Acevedo, 2009). We do not know a lot about the impact of teachers' CPD on students' learning outcomes, but what we do know suggests two important things: that the right kinds of CPD can produce big benefits for learners, and that most of the CPD undertaken by teachers is not of this kind (Joyce and Showers, 2002; Yoon et al 2007; . Wei et al, 2009; Cordingley and Bell, 2012). This evidence is summarised in Figure 9.

As Wiliam (2007) has written,

> *Knowing that* is different from *knowing how*. But in the model of learning that dominates teacher professional development (as well as most formal education), we assume that if we teach the *knowing that*, then the *knowing how* will follow. We assemble teachers in rooms and bring in experts to explain what needs to change—and then we're disappointed when such events have little or no effect on teachers' practice. This professional development model assumes that what teachers lack is knowledge. For the most part, this is simply not the case. The last 30 years have shown conclusively that you can change teachers' thinking about something without changing what those teachers do in classrooms.

Figure 9: What kind of CPD helps learners?

What kind of CPD helps learners?
It should be …

1. **Intense**: at least 15 contact hours, preferably 50
2. **Sustained**: over at least two terms
3. **Content focused**: on teachers' knowledge of subject content & how students learn it
4. **Active**: opportunities to try it out & discuss
5. **Supported**: external feedback and networks to improve and sustain
6. **Evidence based**: promotes strategies supported by robust evaluation evidence

Most of the high-impact strategies from the *Toolkit* (Figure 7, p10, above) are hard to implement; using them effectively is a skill that must be learned. Ironically, teachers are well-placed to know what kinds of practices are likely to be required for learning hard skills. For example, if we want children to learn place value, persuasive writing, music composition or balancing chemical equations, we might start by explaining what they have to do, but then also demonstrating it, getting them to do it (with a fair amount of structure and support to begin with, then gradually reduced), providing feedback to correct and reinforce, getting them to practise, practise, practise until it is really secure, and all the while assessing their performance, reviewing and intervening as necessary. By contrast, if we want teachers to learn hard things like using formative assessment, assertive discipline or how to teach algebra, we often don't get beyond just explaining what they have to do. For a profession that is so dedicated to learning for others, teachers seem to take little care over their own learning.

For many teachers, the kind of CPD described in Figure 9 is rarely offered or experienced. Even if such training were offered, many school leaders would have serious concerns about allowing their teachers to spend so much time out of the classroom, and their CPD budgets would probably not allow it anyway. Moving to a model where CPD like this is routine would require some big culture shifts: a more general recognition that teaching well is hard and needs to be learned; an expectation that teacher CPD should lead to benefits for learners; and an emphasis on evaluating the impact of CPD on learner outcomes.

This change might also require additional investment in CPD; but, on the other hand, it might not. For example, a realistic choice for a school for spending its Pupil Premium could be between (a) reducing classes from 30 to 24 or (b) teaching classes of 30 for four days a week and using the fifth day for CPD.

My second recommendation is therefore: **Invest in effective professional development**

3. Evaluate teaching quality

If we are to capitalise on the benefits of professional development, then we need both to be able to target that development at areas of need and to be able to evaluate its impact. Unless we have sophisticated systems in place for evaluating teaching quality, both are hard to do.

I have already argued that existing approaches to identifying effective schools are problematic; the same arguments apply to attempts to identify effective teachers. Despite that, there has been a growth of interest recently in trying to do this, mostly from the US, mostly based on students' annual gains (on tests of variable quality) using econometric analyses, much of it driven by the desire to create reward structures for effective teachers or 'deselect' the ineffective (eg Hanushek, 2011). Underlying this work is an emerging consensus that it is not largely schools that make a difference to learning, but teachers – a finding which has been evident from the early days of CEM monitoring systems (Coe and Fitz-Gibbon, 1998).

Some of the claims being made by economists are astonishing. For example, Hanushek (2011) uses estimates of the impact that an effective teacher has on the lifetime earnings of their students as a measure of their value. By this metric the difference between a slightly above average teacher (60th percentile) and an average teacher (50th percentile) for a class of 30 is worth $160k for each class they teach for one year. For a really effective teacher (90th percentile) the difference is more than $800k per class. This gives a whole new meaning to the term 'value-added'.

All the concerns about measuring 'effectiveness' that I have already voiced are still relevant. But that does not mean that a constructive evaluation (including self-evaluation) of teaching quality should not be informed by – among other things – evidence from test score gains. The Gates Foundation's Measures of Effective Teaching (MET) Project has recently made important progress in demonstrating the benefits of using multiple measures of effectiveness (see http://www.metproject.org/reports.php). Their three main sources of evidence (test score gains, student ratings and classroom observation) are shown, if done well, to contribute to a single construct of effectiveness, but also to complement each other with each providing a unique contribution. This project has also significantly advanced our understandings of the legitimacy of interpreting 'value-added' as a causal teacher effect, through the use of random allocation (Kane et al, 2013), and of the requirements and limitations of classroom observation (Ho and Kane, 2013).

My recommendation for more evaluation may be seen as a sales pitch for CEM systems. Of course it is that: I am proud to be Director of CEM, I believe passionately in what we do and I want all schools to benefit from it. If teachers see the benefits of using high-quality assessments to monitor their students' progress, of receiving accessible analyses of their exam performance that make fair comparisons with the achievements of similar students in other schools, of being part of a distributed research network that supports them in interpreting and acting on their data, then of course I am happy to enrol them in our systems. But I have always said that if they find those same benefits elsewhere then I am just as happy. What matters is the effective use of good assessment data to inform practice, not who it comes from.

There is a wide range of research evidence about positive impacts of giving teachers feedback on their performance (e.g. Coe, 1998, 2002; Hattie & Timperley, 2007; Hattie, 2009). CEM systems have always focussed on feeding back a range of data analyses to the teachers and school leaders who are in a position both to interpret them in context, and to do something about them if appropriate. We have never claimed that data alone can tell you who is a good teacher or a good school; those are judgements that may properly be supported by appropriate data, but not replaced by it. The difference between most of what is happening in the US under the heading of teacher evaluation and what CEM has been doing for 30 years is that we are interested in evaluation for professional development, not for accountability.

My third recommendation is: **Use multiple sources of validated evidence to support diagnostic and constructive evaluation of teacher quality**.

4. Evaluate impact of changes

My final recommendation is related to the previous one in that it is also focused on evaluation, but this one relates to evaluating the impact of interventions. In some ways this is the most important of all four. Without evaluation, the previous three are just yet more plausible suggestions, with much the same provenance as all the other changes that I claim have not worked.

The message of this talk is not 'What we have done so far has not worked, so let's try some different things,' but 'What we have done so far has not worked, so let's try some different things *and this time evaluate properly whether they work or not*.'

Many educators are lovers of novelty; it is a great strength and a weakness. We invest huge effort and cost in implementing new ideas, and it is likely that some of them bring genuine improvement. However, it is also likely that some – perhaps just as many – lead to deterioration[6]. Many, of course, make no real difference at all. And in most cases we will not know which are which.

This should not depress us; it simply tells us that we cannot judge whether something works without evaluating it properly. We often think we can, but the evidence shows we are wrong.

There are some important recent national developments that are really encouraging here. The work of the Educational Endowment Foundation has really raised the priority and level of thinking about high-quality evaluation in education across England. Following the Goldacre review[7], large-scale randomised controlled trials have been launched by the Department for Education[8] and the National College for Teaching and Leadership[9].

Evaluation is not simple to do, so just saying 'we should evaluate' is not the same as building the capacity and culture required to make it happen. We must not underestimate the challenge of doing this, and must plan and resource that development carefully. One contribution to this is the Educational Endowment Foundation's DIY Evaluation Guide (Coe et al, 2013), designed to support teachers in evaluating their own small-scale changes.

My fourth and final recommendation is: **Whenever we make a change we must try to evaluate its impact as robustly as we can**.

Conclusions

I have claimed that educational standards in England have not risen, attempts to improve education have largely failed and even when we think we can pick out the best schools, we "know so many things that just ain't so".

Actually it doesn't really matter whether or not you accept this analysis; either way we should look for the best strategies to bring real improvement. Even if hope is not rational or evidence-based, we need to hold on to it. Education is far too important to give up on.

I have made four suggestions. The first three are just suggestions; I'll be happy to concede them if others have better suggestions. But the fourth is non-negotiable: the fourth is the one that distinguishes what I am recommending from what we have done before. Education has existed in a pre-scientific world, where good measurement of anything important is rare and evaluation is done badly or not at all. It is time we established a more scientific approach.

[6] Two reviews from medicine suggest that even for therapies that get as far as being subjected to randomised clinical trial, the innovative treatment proves to be better than the existing in only about half of all cases (Gilbert et al, 1977; Soares et al, 2005)
[7] https://www.gov.uk/government/publications/department-for-education-analytical-review
[8] https://www.gov.uk/government/news/new-randomised-controlled-trials-will-drive-forward-evidence-based-research
[9] http://www.education.gov.uk/nationalcollege/testandlearn

References

Bill and Melinda Gates Foundation (2012) 'Ensuring Fair and Reliable Measures of Effective Teaching'. Measures of Effective Teaching (MET) Project research paper, January 2012. Avaliable at http://www.metproject.org/reports.php

Black, P. and Wiliam, D. (1998) Inside the Black Box. London: King's College.

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal*, 29(5), 623-637.

Brown, M. (1998) 'The tyranny of the international horse race'. In R. Slee and S. Weiner (with S. Tomlinson) *School Effectiveness for Whom*?. London: Falmer.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (No. w17699). National Bureau of Economic Research.

Coe, R. (1998) 'Can Feedback Improve Teaching?' *Research Papers in Education*, 13, 1, 43-66.

Coe, R. (2002) 'Evidence on the Role and Impact of Performance Feedback in Schools' in A.J.Visscher and R. Coe (eds.) *School Improvement Through Performance Feedback*. Rotterdam: Swets and Zeitlinger (23pp).

Coe, R. (2007) Changes in standards at GCSE and A-Level: Evidence from ALIS and YELLIS. Report for the Office of National Statistics, April 2007. Curriculum, Evaluation and Management Centre, Durham University. Available at http://www.cemcentre.org/attachments/ONS%20report%20on%20changes%20at%20GCSE%20and%20A-level.pdf

Coe, R. and Tymms, P. (2008) 'Summary of Research on Changes in Educational Standards in the UK', in M. Harris, *Education Briefing Book 2008: IoD Policy Paper*. London: Institute of Directors, August 2008.

Coe, R., Fitz-Gibbon, C.T. and Tymms, P. (2000) *Promoting Evidence-Based Education: The Role of Practitioners*. Roundtable presented at the British Educational Research Association annual conference, Cardiff, September 2000. [Available from http://www.leeds.ac.uk/educol/]

Coe, R., Kime, S., Nevill, C. and Coleman, R. (2013) 'The DIY Evaluation Guide'. London: Education Endowment Foundation. [Available at http://educationendowmentfoundation.org.uk/library/diy-evaluation-guide]

Coe, R.J. and Fitz-Gibbon, C.T. (1998) 'School effectiveness research: criticisms and recommendations'. *Oxford Review of Education*, 24, 4, 421-438

Cordingley, P. & Bell, M. (2012) Understanding What Enables High Quality Professional Learning: A report on the research evidence. Centre for the Use of Research Evidence in Education (CUREE); Pearson School Improvement http://www.pearsonschoolmodel.co.uk/wp-content/uploads/2011/09/CUREE-Report.pdf

Davies, P. (1999). What is evidence-based education?. *British Journal of Educational Studies*, 47(2), 108-121.

Dumay, X., Coe, R., and Anumendem, D. (in press, 2013) 'Stability over time of different methods of estimating school performance'. *School Effectiveness and School Improvement*, vol , no , pp . http://www.tandfonline.com/doi/full/10.1080/09243453.2012.759599

Fitz-Gibbon, C.T. (1996) *Monitoring Education: indicators, quality and effectiveness*. London: Cassell.

Gilbert, J.P., McPeek, B. and Mosteller, F. (1977) 'Statistics and ethics in surgery and anesthesia'. *Science* 198(4318):684–689

Gorard, S. (2010) Serious doubts about school effectiveness, *British Educational Research Journal*, 36, 36: 5, 745—766.

Gorard, S. and Cook, T. (2007) Where does good evidence come from? *International Journal of Research and Method in Education*. Vol. 30, No. 3 (2007): 307-323.

Hanushek E.A. & Woessmann L. (2010) The High Cost of Low Educational Performance: The Long-run Economic Impact of Improving PISA Outcomes. OECD Programme for International Student Assessment. http://www.oecd.org/dataoecd/11/28/44417824.pdf

Hanushek, E.A. (2011) The economic value of higher teacher quality. *Economics of Education Review* 30 466–479

Hattie, J. & Timperley, H. (2007) 'The Power of Feedback', *Review of Educational Research*, 77, 1, pp. 81-112.

Hattie, J. (2003) Teachers Make a Difference: What is the research evidence? Australian Council for Educational Research, October 2003.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E., & Coe, R. (2013). The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit. London: Education Endowment Foundation. [Available at http://www.educationendowmentfoundation.org.uk/toolkit]

Hill, P.W. (1998). Shaking the foundations: Research driven school reform. *School Effectiveness and School Improvement*, 9, 419-436

Ho, A.D. and Kane, T. J. (2013). 'The Reliability of Classroom Observations by School Personnel'. Research Paper. MET Project. Bill & Melinda Gates Foundation. Available at http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf

Hodgen, J., Brown, M., Coe, R., & Küchemann, D. E. (2012). Why are educational standards so resistant to reform? An examination of school mathematics in England. Paper presented at the 2012 Annual Conference of the American Educational Research Association (AERA), Vancouver.

Hodgen, J., Kuchemann, D., Brown, M. and Coe, R. (2009) 'Children's understandings of algebra 30 years on'. Research in Mathematics Education, 11, 2, 193-194.

Hodgen, J., Küchemann, D., Brown, M., & Coe, R. (2010). Multiplicative reasoning, ratio and decimals: A 30 year comparison of lower secondary students' understandings. In M. F. Pinto & T. F. Kawaski (Eds.), Proceedings of the 34th Conference of the International Group of the Psychology of Mathematics Education (Vol. 3, pp. 89-96). Belo Horizonte, Brazil.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

Jerrim J. (2011) England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline? Institute of Education, DoQSS Working Paper No. 11-09, December 2011. Available at http://www.ioe.ac.uk/Study_Departments/J_Jerrim_qsswp1109.pdf

Joyce, B.R., and Showers, B. (2002) 'Student Achievement through Staff Development' 3rd ed. ASCD www.ascd.org [See also an excellent practical summary at www.geoffpetty.com/downloads/WORD/JoyceNshowers.doc]

Kane, M. T. (2006) Validation. In Robert L. Brennan (Ed.), *Educational Measurement*, 4th Edition. Westport, CT: American Council on Education and Praeger Publishers.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. Bill & Melinda Gates Foundation. Available at http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kukla-Acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. Economics of Education Review, 28(1), 49-57.

Luyten, H., Visscher, A. and Witziers, B. (2005) School Effectiveness Research: From a review of the criticism to recommendations for further development. *School Effectiveness and School Improvement*, 16, 3, 249-279.

Mihaly, K., McCaffrey, D.F., Staiger, D.O. and Lockwood, J. R. (2013) 'A Composite Estimator of Effective Teaching'. Measures of Effective Teaching (MET) Project research paper. Avaliable at http://www.metproject.org/reports.php

Nuthall, G. (2004). Relating classroom teaching to student learning: A critical analysis of why research has failed to bridge the theory-practice gap. Harvard Educational Review, 74(3), 273-306.

Nuthall, G. (2005). The cultural myths and realities of classroom teaching and learning: A personal journey. The Teachers College Record, 107(5), 895-934.

Nuthall, G. (2007). *The hidden lives of learners*. NZCER Press.

OECD (2010), PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I). http://dx.doi.org/10.1787/9789264091450-en

Ouston, J. (1999) 'School effectiveness and school improvement: Critique of a movement', in T. Bush, R. Bolam, R Glatter and P Ribbins (ed.s) *Educational Management: Redefining theory, policy and practice*. London: Paul Chapman.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.

Robinson, V. M. J., Lloyd, C. A. and Rowe, K. J. (2008) The Impact of Leadership on Student Outcomes: An Analysis of the Differential Effects of Leadership Types. *Educational Administration Quarterly* 2008; 44; 635.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. The American Economic Review, 94(2), 247-252.

SCAA (School Curriculum and Assessment Authority) (1996) GCSE Results Analysis: an analysis of the 1995 GCSE results and trends over time. London: SCAA.

Scheerens J. (2000) 'Improving School Effectiveness' UNESCO http://unesdoc.unesco.org/images/0012/001224/122424E.pdf

Scheerens, J. (ed.) (2012) *School leadership effects revisited: review and meta-analysis of empirical studies*. Dordrecht: Springer.

Scheerens, J., Bosker, R. J. and Creemers, B. P. M.(2001) 'Time for Self-Criticism: on the Viability of School Effectiveness Research', *School Effectiveness and School Improvement*,12:1,131 — 157

Shadish, W., Cook, T. D., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Miflin.

Shayer, M., Ginsburg, D. and Coe, R. (2007) 'Thirty Years on—a large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975-2003'. *British Journal of Educational Psychology*, 77, 1, 25-41.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, 31(4), 500-506.

Soares, H. P., Kumar, A., Daniels, S., Swann, S., Cantor, A., Hozo, I., ... & Djulbegovic, B. (2005). Evaluation of New Treatments in Radiation Oncology Are They Better Than Standard Treatments?. *Journal of the American Medical Association*, 293(8), 970-978.

Stewart, W. (2012) 'Gove accused of building on shaky Pisa foundations'. *Times Educational Supplement*, 2 November 2012. http://www.tes.co.uk/article.aspx?storycode=6298801

Teacher Development Trust (2013) "The new theory of evolution" http://www.teacherdevelopmenttrust.org/the-new-theory-of-evolution/

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, 49(3), 316-328.

Tymms, P. and Merrell, C. (2007) Standards and Quality in English Primary Schools Over Time: the national evidence (Primary Review Research Survey 4/1), Cambridge: University of Cambridge Faculty of Education.

Tymms, P. B., Merrell, C., & Coe, R. (2008). Educational policies and randomized controlled trials. *Psychology of Education Review*, 32 (2), pp3-7 & 26-9.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad. Dallas, TX. National Staff Development Council. http://www.becker.k12.mn.us/sites/beckerschools/files/departments/2012/NSDCstudytechnicalreport2009_0.pdf

Wiliam, D. (2009). Assessment for learning: why, what and how? (Inaugural Professorial Lecture) London: Institute of Education, University of London.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033.pdf