Standards in English schools: changes since 1997 and the impact of government policies and initiatives.

A report for the Sunday Times

Professor Peter Tymms, Dr. Robert Coe and Dr. Christine Merrell

CEM Centre, University of Durham,

April 2005

The opinions expressed in this report are the personal opinions of the authors and do not present the official position of Durham University.

Introduction

This report for the Sunday Times looks at the attainments of pupils in England since the Labour Party came to office in 1997 and where possible earlier. It seeks to estimate the impact that recent changes have had on the attainments of pupils. It does so by looking at official test data and at other data where it is available.

The report side-steps some of the academic discussions surrounding words such as “attainment” and “standards”. “Standards” can sometimes refer to the attainments of pupils, but it is more properly used to refer to the benchmarks used by examination authorities, thereby causing some confusion. We will use the word to mean “the attainments of pupils”. We also largely side-step the issue as to whether the exam results truly reflect the performance of children, whether they are useful measures and if they do increase, if they represent genuine gains in educational terms. In other words, we will concentrate only on the official statistics and whether they have indicated rises in attainments by pupils. In order to do this we will look at independent test data. Then we look at the magnitude of the gains, and whether they represent value for money.

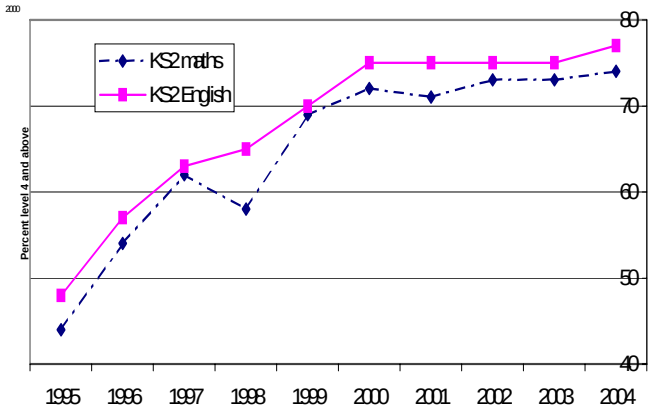
Changes in the attainments of pupils in primary schools

This section of the report will look at the attainment of pupils in primary schools in England from a number of perspectives. First, the official data, the end of Key Stage 2 results will be described. Secondly, independent reports and relevant data that exist, both official and independently collected, will be itemised and described. Next, the results will be brought together and summarised together with a description of the findings of the February Statistics Commission Report. Finally, conclusions will be drawn about how the attainments of pupils in English primary schools have changed over the last decade.

Official data

The chart below and the following table show the percentage of children in England who have attained a Level 4 or higher in the statutory end of Key Stage 2 results for mathematics and English. The results are shown from 1995 to 2004 and they indicate for English a steady rise between 1995 and 2000 and more or less constant results thereafter. The mathematics scores rise steadily again from 1995 to 2000 and then modest rises thereafter, with an anomalous drop in 1998.

Percent of pupils gaining a Level 4 or above at the end of Key Stage 2



Year	KS2 maths % Level 4 or higher	KS2 English % Level 4 or higher
1995	44	48
1996	54	57
1997	62	63
1998	58	65
1999	69	70
2000	72	75
2001	71	75
2002	73	75
2003	73	75
2004	74	77

Bearing in mind that the Labour Party came to office in 1997, these data would suggest that the attainment of children in English and mathematics continued to rise during the first three years of their office and remained fairly constant thereafter. The rises were quite dramatic between 1997 and 2000. For example, in English the rise from about 63% to about 75% is clearly a large increase and if it were genuinely to reflect the attainment of pupils it would represent a considerable impact on children and their ability to read and write. The question arises as to whether these percentages can be seen as reflecting a real change in attainment or if there is some other explanation.

Reports and data sources

There have been a number of independent reports which have looked at standards over time or which are relevant to the issue of attainment over time which are listed below together with various sources of official and independently collected data. In addition to the official testing there are 13 different sources of assessment results which amount to the testing of about one third of a million children.

1. The end of Key Stage 2 results from English and mathematics as in the diagram and table above.
2. The end of Key Stage 3 results from English and mathematics. This is important because it shows what the children were doing in secondary school three years later. Of course, the interpretation of such data is open to question but one might expect, genuine gains at Key Stage 2 to be followed by genuine gains at Key Stage 3.
3. Dr. Julie Davis and Ivy Brember of Leeds University have produced a series of papers based on their data collection in the same five randomly chosen schools from one LEA since 1989. The latest reports results up to 1998.
4. Professor Margaret Brown et al (2003) of King's College London studied two cohorts of children as they progressed through primary school as part of a Leverhulme funded study looking at the maths attainment of children.
5. Mary Hilton of Cambridge University looked carefully at the English statutory tests for each year from 1998 to 2000. She found that "the reading tests were progressively easier for the children to answer. … because the number of higher-order reading skills … has decreased each year".
6. A report by Massey et al for QCA used a retest approach to check on the standard setting of the Key Stage 2 results from 1996, 1999 and 2000. This careful large-scale study costing £300,000 and commissioned by QCA investigated the standards applied to English statutory tests from 1996 and 1999[1]. The tests were administered to equivalent samples of children in Northern Ireland at the same time. The proportion of children gaining level 4 and above using the 1996 and 1999 procedures were then compared. If the standards set at Key Stage 2 were equivalent then the results in Northern Ireland from the tests from 1996 and 1999 should have produced the same results.
   The report also reported data from 6 different Local Education Authorities

[1] Some of the 2000 tests were also investigated in and extension study.

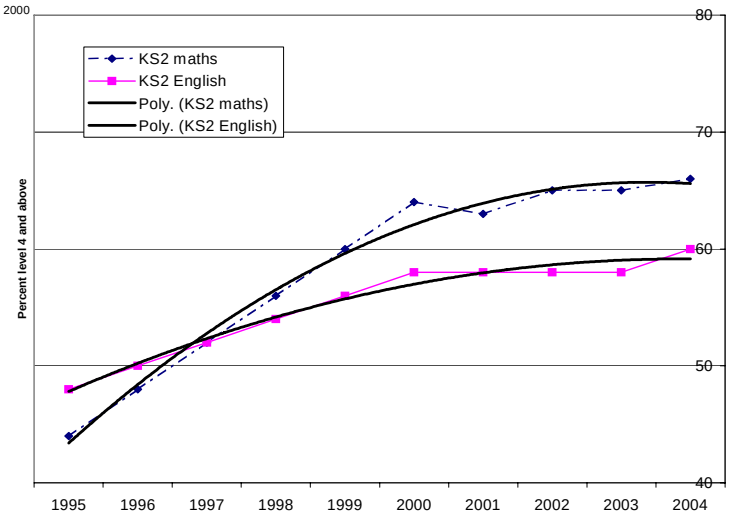
with results on the same tests for reading, maths and writing (not from all LEAs) between 1996 and 1999.

7. A study commissioned by QCA/DfES and reported by NFER collected data from pupils in Years 4 and 5 repeatedly from 1998, 1999 and 2000.
8. Professor Tymms and Professor Fitz-Gibbon of Durham University reported results generated by tests run from the Curriculum Evaluation and management (CEM) Centre in a chapter published in 2001.
9. The Performance Indicators in Primary Schools (PIPS) project, from the CEM Centre at Durham University, has used reading and mathematics tests in all years of primary education in thousands of schools for a number of years. The most relevant are from Year 6 and year 4 where the tests every year gives data going from 1997 up to 2004.
10. The Middle Years Information System (MidYIS), from the CEM Centre at Durham University, is a monitoring project for secondary schools which has a very widely used baseline assessment for secondary schools. This project generated maths test data from pupils in Year 7 from 1999 onwards.
11. Professor Tymms brought much of the above results together in a paper published earlier this year in the British Educational Research Journal.
12. The most recent and perhaps the most important report to have appeared is the summary by the Statistics Commission. Their report, Number 23, February 2005, dealt specifically with the measuring of standards in English primary schools.
13. Since then reading test data from a seventh LEA has become available. They used the same test with pupils in Year 7 for many years.

Summarising the results from the reports and datasets.

The amalgamation of all these data is quite complex and subject to careful analysis which there is not space to report here. Instead reliance is placed on the 2004 paper by Tymms who came to the conclusion that the massive rises between 1999 and 2000 in the end of Key Stage 2 Assessments over-estimated the actual rise in pupils' results. One point to note is that there is a slightly different pattern for English and mathematics and that the English data are in fact made up of writing and reading scores. So little information is available on writing as a separate entity that no comment is made on it here. For reading the picture is clear. The Massey Report, which looked at End of Key Stage 2 assessments from 1996 to 1999, called the rise in reading "illusory". The rise in maths seemed to be a little larger but nevertheless small. It is possible using all the information to estimate the true rises and it would seem that over the years the percentage of children achieving Level 4 or above in English should have risen from 48% to 58%, not the figure of 75% recorded by QCA. In mathematics, the percentage of children achieving a Level 4 or above should have risen from 44% to 64%, not to the officially reported figure of 72% with the percentage continuing to rise slightly after the year 2000. A truer picture of the changes in results is estimated in the chart and table below:

Percent gaining a Level 4 or above (Key Stage 2) – a corrected simplified picture



Year	KS2 maths % Level 4 or above	KS2 English % Level 4 or above
1995	44	48
1996	48	50
1997	52	52
1998	56	54
1999	60	56
2000	64	58
2001	63	58
2002	65	58
2003	65	58
2004	66	60

There was no indication of a particular kick after 1997 after the Labour party gained office and the observation that the official Key Stage 2 results remained more or less constant for English since 2000 are confirmed by the independent data.

The evidence was reviewed very recently by the Statistics Commission and in their February report they stated:

"*The Commission believes that it has been established that (a) the improvement in KS2 test scores between 1995 and 2000 substantially overstates the improvement in standards in English primary schools over that period, but (b) there was nevertheless some rise in standards*."

"*Ministers, and others who may want to use the test scores in a policy context, need to be made fully aware of any caveats about their interpretation. As Tymms's article demonstrates, the sharp rise in KS2 scores in the latter*

1990s cannot be simply interpreted as a rise in schools performance standards - there are a number of qualifications that need to be made. Yet Government Departments have usually failed to mention any caveats about other possible reasons for rising test scores in their public comments."

Changes in the attainments of pupils in secondary schools

In this section we look at changes in performance since 1997 in national examinations at GCSE and GCE A Level.

The annual publication of A level and GCSE results has, in recent years, been accompanied by claims of falling standards as each year more people achieve higher grades. Despite this concern, and a number of attempts to address the issue, the evidence about whether standards have actually fallen is unclear. This paper presents new evidence which matches the performance of a sample of students in public examinations with their performance on a stable aptitude test.

We look first at official data showing the rises in attainment in public examinations over the last thirty years and, in particular, at the scale of changes over the last seven years. Second, we examine changes in performance on a fixed test (the Yellis test) taken by large numbers of GCSE candidates since 1995. Third, we look at how students of the same ability (as measured by the Yellis test) have performed in GCSEs over the same time period. Fourth, we look at how students of the same ability (as measured by the TDA test used Alis) have performed at Advanced level over a period of some 17 years.

Official data

In terms of raw results, there have been significant rises in performance at both GCSE and A Level, both before and after 1997. These are shown in Figures 1 and 2 respectively.

The question of whether these rises represent genuine improvements in learning and attainment, or just a gradual lowering of standards, is a controversial one.

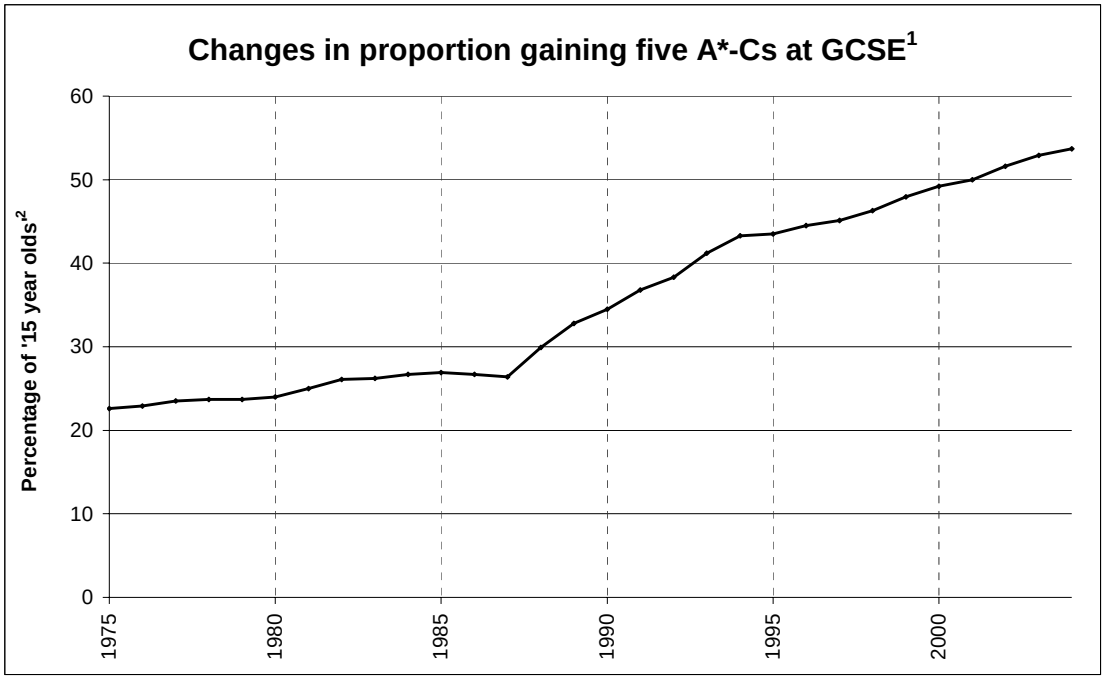


Figure 1
Notes 1& 2 Prior to 1988, percentage is for A-C at O Level and grade 1 CSE

At GCSE, the official statistics show that the percentage rose slowly, increasing by just a third of a percentage point each year until the change from O Level/CSE to GCSE in 1988, when a period of much faster increase began. Between 1988 and 1994 the average annual improvement was almost two and a half percentage points. From 1995 the rate of growth slowed and has remained roughly steady since then until the present day increasing at a rate of one percentage point per year.

There is no clear change in the pattern after 1997, so it is not obvious that policies implemented since then have altered the rate of improvement. For comparison with other changes over the same time period (see below), the change in this indicator can be expressed as a standardised Effect Size[2]. In this metric, the Effect Size for the change between 1997 and 2004 is 0.2.

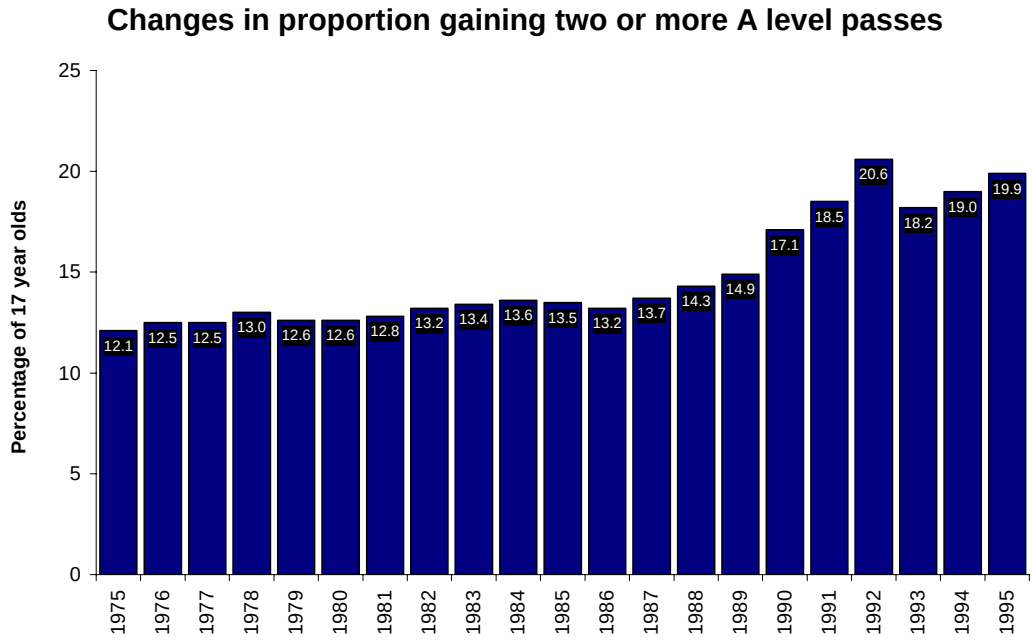


Figure 2: Proportion gaining two of more passes at A level between 1975 and 1995.

Evidence from independent data: ALIS and YELLIS

About ALIS and YELLIS

ALIS (the A Level Information System) began in 1983 as a system for helping schools to compare the progress their students have made with that of students in other schools. Currently over 1400 schools and colleges participate in the project, which processes about half of the A levels taken in the UK. Schools receive value added analysis for individual subject entries, based on simple residual gains when A level scores are regressed on average GCSE scores, as well as data on a range of student attitudes and perceptions. An optional part of the scheme is the Test of

[2] The term "Effect Size" is widely used in educational research as it can be applied universally. If an Effect Size is 0.2 or less it is regarded as small. If it is 0.8 or greater it is regarded as big.

Developed Abilities (TDA), which is offered free of charge to participants in ALIS, should they wish to have an additional base-line measure from which to calculate value added.

YELLIS (Year 11 Information System) began in 1994 and now analyses the GCSE results of about 1300 schools. YELLIS uses its own test, taken by students in year 10 or year 11, as a base-line from which to calculate value added. The YELLIS test comprises two sections, mathematics and vocabulary.

For the purposes of this analysis, both projects provide two kinds of data. The first is a set of scores on a measure of general academic ability that has remained constant over a number of years. This avoids one of the problems of trying to compare GCSE or GCE grades over time which is that the examination is different each year. Inevitably, though, it raises further problems of whether the test is appropriate to use as a measure of attainment and, if so, whether its appropriateness remains constant over the time period in question. These are controversial matters and the subject of dispute by academics and examiners.

The second is data on the relationship between those ability test scores and subsequent performance in national examinations. Knowing the ability of students who have taken a particular exam in a particular year enables us to compare their achievement with other students of similar ability in other years.

Changes in performance on the Yellis test

The Yellis test average scores for all students who took the test in either Year 10 or Year 11 between 1995 and 2004 are shown in Figure 3. Given that the numbers are quite small in the early years of Yellis, we should interpret the results for 1995 and 1996 with caution. Hence the baseline from the time before the election of the New Labour government in 1997 is less reliable than after that time.

Since that time, however, there appears to have been a small but steady rise in the average scores of both Y10 and Y11 students. Between 1997 and 2004 the increases for these groups amount to effect sizes of 0.2 and 0.4 respectively. The fact that the scores are increasing suggests that students' mathematical and verbal abilities have improved slightly over that period, though the increase on this scale over seven years is quite small. Nevertheless, this is certainly consistent with the claim that standards of educational attainment have genuinely risen.

In addition, the fact that the gap between Y10 and Y11 scores is widening could be taken as further confirmation of this claim, since the gap effectively represents the progress made during Year 10. Of course, this interpretation depends on the fact that both groups of students are representative of their respective year groups.

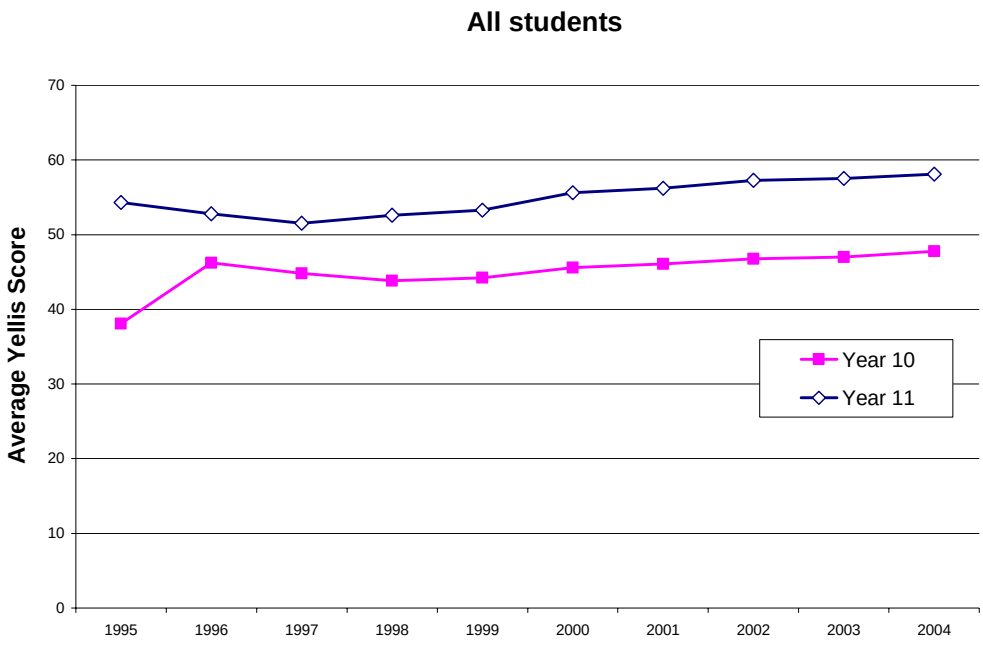


Figure 3

Changes at GCSE in relation to the Yellis test

Against the background of continuing rises in national GCSE performance (Figure 1) and a steady rise in Yellis test scores (Figure 3), the question remains about whether students with the same Yellis score achieve more or less at GCSE over time.

For this analysis, we take the relationship between the Yellis test taken in Y10 and GCSE grades achieved in examinations about 20 months later. The numbers of students in this group are much larger than for the Y11 group for all but the first few years of Yellis. For simplicity we take a typical Yellis test score of 45 as our reference point. The comparison therefore shows the average grade achieved by students who scored 45 on the test in Y10, and hence controls for the ability of the student.

Overall, the averages of all GCSE grades achieved each year by these students with 45 on the Yellis test are shown in Figure 4. Twenty-six separate GCSE subjects were analysed by Yellis throughout this period. GCSE grades are coded numerically as A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0. Again if we discount 1995 and 1996 where the numbers are smaller and the membership of Yellis less stable, we see a very tiny rise in grades achieved, though this is rather dependent on which year one takes as the start and finish of a trend, since the year-to-year fluctuation is about as big as any overall change. In fact the average rise for the period 1997 – 2004 is only about $1/60^{th}$ of a grade per year, so for all practical purposes, the line is flat. Overall, therefore, we can say that for GCSE subjects as a whole, students of comparable ability achieved the same grades regardless of which year they took their exams during this period.

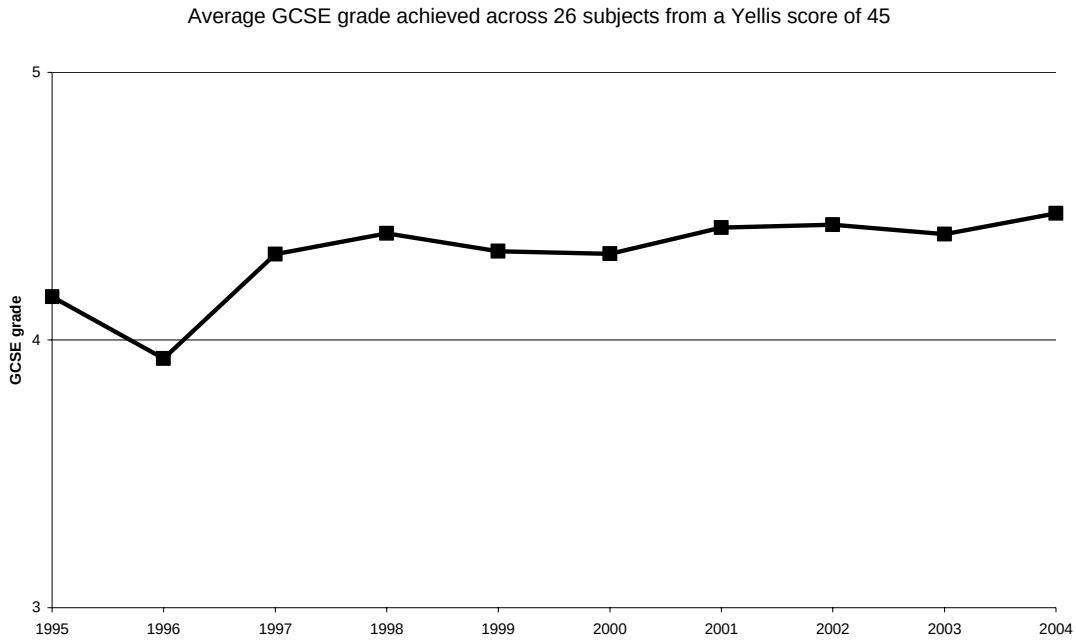


Figure 4

If we look separately at the five subjects with the largest number of entries (double science, English, French, history and maths), we see a somewhat mixed picture (Figure 5).

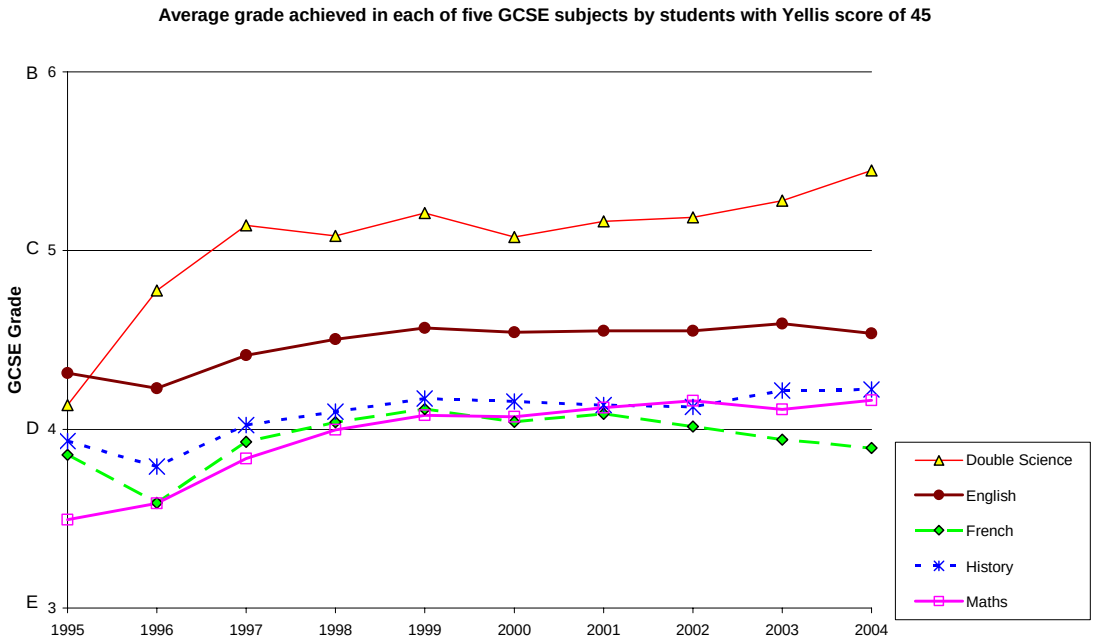


Figure 5

Science and maths have both risen fairly steadily; students in 2004 achieved about a third of a grade higher than those of similar ability had done in 1997. In maths the main part of this rise occurred at the beginning of this period, while in science it

was mainly towards the end. English and history GCSE rose a little between 1997 and 1999 but then levelled off. Performance by students of matched ability also rose until 1999 in French, but then fell again, ending the period in 2004 lower than it began in 1997.

For those subjects where there has been a rise in the performance of students of the same ability (e.g. science and maths), there are a number of possible explanations. One is that teaching has improved, leading to higher achievement by students of the same ability. Another is that changes in the examination grading have altered the standard and made the same grades easier to achieve.

Changes at A-Level in relation to the TDA

Data showing the relationship between TDA scores and performance in a range of A levels are available from 1988. The six subjects with the largest entries in ALIS are analysed here: Biology, English, French, Geography, History and Mathematics.

There are some problems with making comparisons over such a period. Inevitably syllabuses change and it is not always straightforward to decide whether what is being compared is quite the same. In particular, a subject like mathematics includes a number of different syllabuses but excludes others. For example, until 2001 modular syllabuses were not included under this heading; in 2002, with the start of *Curriculum 2000*, all A level syllabuses effectively became modular. English is also somewhat problematic, with different syllabuses in literature and language, or mixtures of the two.

A further problem is that the TDA, used since 1988, was modified slightly in 2000 in order to improve its predictions. The original test was known as the International Test of Developed Abilities and the new test was known more simply as the Test of Developed Abilities (TDA). This change has resulted in a discontinuity between (I)TDA scores for students taking A levels up to 2001 and those taking them from 2002 onwards. Although the new test is slightly easier than the old one, in the analyses presented here a correction has been applied to equate scores from the two tests.

Changes in performance on the Test of Developed Abilities (TDA)

Changes in the TDA scores of candidates in the six subjects are shown in Figure 6.

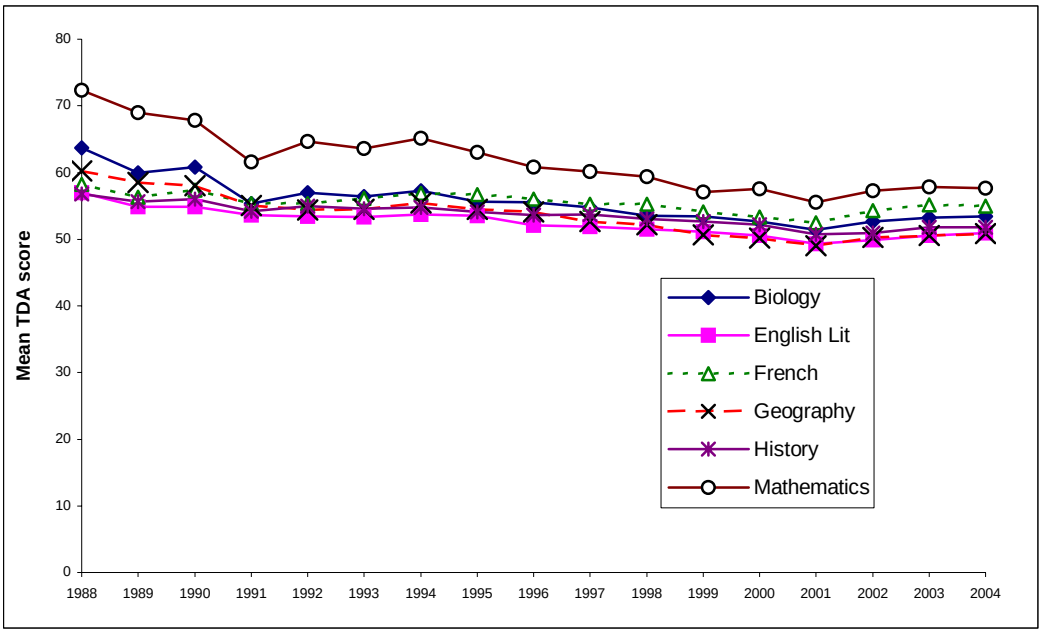


Figure 6 Mean TDA score for students taking a range of A level subjects.

The data shows that there has been a decline in the TDA scores of the candidates from 1988 to 2001 and a possible slight increase thereafter. This was more noticeable for Mathematics.

Figure 7 shows the A level grades expected of candidates having the same TDA score each year since 1988 for the six subjects. From 1988 until 2004 the achievement levels have risen by about 1½ grades across all subjects on average. Exceptionally, from 1988 the rise appears to be about 3 grades for Mathematics. This could be due to this severely graded subject being brought more into line with other subjects. On the surface of it would appear that there has been a significant increase in the grades achieved by students, and there could be much speculation as to why this has arisen. There are those who would argue that as more students take A level a norm referencing component of the grade awarding process will inevitably lead to a small gradual upward drift in grades, with a significant cumulative effect over a period of years. On the other hand examination boards could site the presence of the work of the previous year's candidates at the awarding process, coupled with extensive statistical analysis ensures that standards are being maintained. From a different perspective it could be argued that A level teaching has improved or that syllabuses have been cut back and exams made easier.

It is our view that A levels have generally become more leniently graded through a combination of syllabus change, modularisation and alterations to the exam formats. In many ways that has been a good thing. It has allowed increasing numbers of candidates to access education to higher levels. But it has meant that the very top levels of attainment have been removed from A level.

A level results and TDA scores

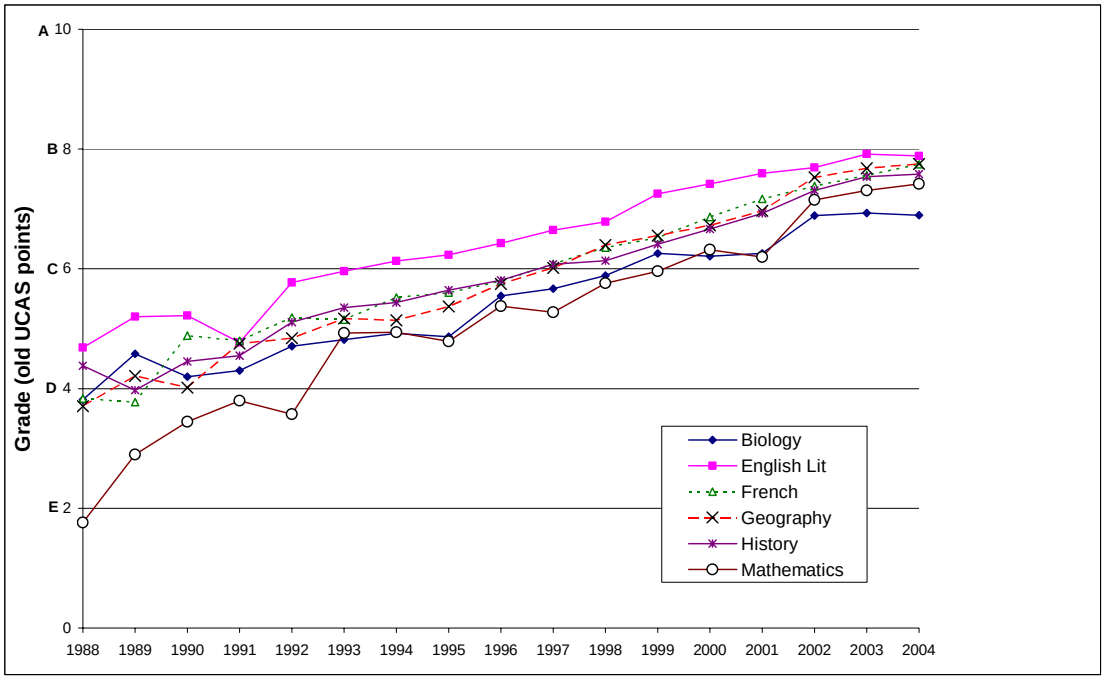


Figure 7: Mean UCCAS grades achieved with a students with a TDA score of 60.

Summary

The picture in secondary schools is complex.

At the end of compulsory schooling GCSE passes have been rising for some time. The rises did not suddenly happen after 1997. Indeed the year-on-year rises from 1987 to 1994 were greater than the rises after 1995, which have been constant ever since.

Independent measures suggest that there has been real and sustained growth in attainment over the last decade at the end of compulsory schooling. But that rise has been small and it would be hard to maintain that the rise occurred solely since 1997.

A level attainment levels are difficult to pin down as syllabuses have changed as has the number of students taking exams, Curriculum 2000 has been introduced and modularisation has become universal. Nevertheless, it is clear that a higher proportion of candidates are getting higher grades. Careful analysis of grades in relation to independently collected test data shows that higher grades are being awarded to pupils of the same ability. This drift has been apparent for at least 15 years and has not suddenly appeared. The cumulative effect is large but the year-on-year change is small.

It is our opinion, in the face of competing explanations, that A level has become more leniently graded over the years. In some ways this has been a good thing in that it has allowed more students to access more education. But it has been at the expense of very high attainment levels.

International Studies of Students' Achievements and Attitudes

Quantitative comparative studies of education systems and practices across the world can usefully inform the direction of research and policy in individual countries and enable countries to monitor trends over time. The data that they provide begin to answer questions such as: which countries' education systems are the most effective? They also raise questions about the different approaches taken in different countries such as: what age should children to start school? For countries to learn from each other, fair and relevant international comparisons of the attainments and attitudes of students at varying times in their educational careers are necessary. Several large-scale comparative studies have been conducted, some on a regular basis. Three are discussed below.

PIRLS

The Progress in International Reading Literacy Study (PIRLS) was conducted in May 2001 with the support of the International Association for the Evaluation of Educational Achievement (Mullis, Martin, Gonzalez and Kennedy, 2001). It included a sample of 3156 children in Year 5 (average age of 10.2 years) from schools in England. England's rank position of 3rd out of the 35 countries that participated in the study has been the focus of much attention and cited as evidence that the Literacy Strategy, introduced in 1998, is raising standards. Following the release of the results, Education and Skills secretary Charles Clarke said: "We have much to be proud of, and I congratulate our primary schoolchildren for performing so well on the international stage. The fact that our 10-year-olds are reading at a higher level than almost every other country is a credit to them and our education system. It shows that the National Literacy Strategy we set up five years ago to raise standards in primary schools is working." (Labour Party, 2003)

In fact the mean score of English pupils was significantly lower than those in Sweden, who were ranked 1st but not significantly lower than pupils from The Netherlands, ranked 2nd. Looking at countries ranked below England, the mean score of the English pupils was not significantly higher than Bulgaria, ranked 4th but was significantly higher than all other lower-ranked participating countries. Although the average score of pupils in England was high, there remained an issue with the range of scores, which was wider than the international average range of scores, particularly in the lower quartile. Also, despite their achievements in reading, children in the study from England tended to be less keen to read and to be less confident about their ability to read than children in many other countries. While the politicians have celebrated the results of PIRLS as a demonstration of the success of their policy, more detailed research has highlighted issues with the methodology of the study and put into question the validity of its findings.

Clark (2004) raised several issues. Firstly, the cohort of pupils tested would have started school in 1996 and have started to learn the basics of reading before the implementation of the Literacy Strategy. Although England's pupils were slightly younger than the international average they had received 5 years of formal schooling, which was one year more than the international average. *When a factor as important as the number of years of formal schooling ranges from as low as 3 years to as high as 5, the comparability of the results begins to look questionable unless the years of formal schooling have been controlled for.*

Further, the sample of pupils in England was problematic in that it met the guidelines for participation rate only after replacement schools were included and it covered less than 95% of the National Desired Population. From the 150 schools first selected to participate, just 88 agreed. Others had then to be invited. As this procedure continues, the sample of schools that eventually participate is biased towards those who are disposed to such studies and might not therefore be representative of the country as a whole, particularly when surveying attainment and attitudes. Across the participating countries, special schools were excluded from the sample. Within England, pupils were also excluded if they met one of the following conditions: They had a physical disability, statement of special educational needs, referred for multiprofessional assessment, judged by the teacher to be 'temporarily unable to cope with the test conditions' on the day of the test, if English was not their first language and they had been taught in English for less than a year, if in the professional opinion of the teacher, despite having had a year of education in English they still lacked fluency in reading and writing the English Language (Twist, Sainsbury, Woodthorpe and Whetton, 2003). *This ensures that the English sample cannot be fairly compared with other countries*. The effect on the mean reading score of these exclusions is obvious.

In an article submitted for publication, Mary Hilton has discussed in detail many problems with PIRLS including the difficulties associated with trying to eliminate cultural and linguistic bias within the test and questions if it is actually possible and meaningful to produce a unidimensional indicator of reading literacy achievement. She also pointed out that the tests were constructed by a team which included a significant number of test developers from the NFER, who also design the English National Curriculum tests. These English researchers also developed a definition of reading literacy that was the basis of the rationale for the PIRLS test. The English education system has placed emphasis on the particular skills defined within the term 'reading literacy' and pupils are well practiced at completing national tests of a very similar format. They are therefore placed at an advantage compared with pupils in other countries whose curriculum might have a different focus and whose previous experience of assessments was very different to the English one.

Taken together the issues outlined above suggest that PIRLS must be interpreted with caution and, whilst being of interest, it is not robust enough to be considered as evidence of the success of educational policies.

Taking the points discussed above as a warning against over-interpretation of results from international studies, two further studies, the Trends in International Mathematics and Science Study (TIMSS, 2003), and the Programme for International Student Assessment (PISA, 2003) are described below.

PISA

PISA started in 2000 and is conducted on a three-year basis by the Organisation for Economic Co-operation and Development (OECD) (PISA, 2003). The design and implementation of the tests was co-ordinated by the Australian Council for Educational Research. PISA includes assessments of reading, mathematical and scientific literacy. All three domains are assessed in each cycle, with a specific focus on one domain each time. The focus in 2000 was reading. 32 countries participated, including England where 4120 students in Year 11, born in

1984, from 150 schools were selected for assessment. These pupils had been through the English education system after the introduction of the National Curriculum but before the Literacy Strategy. England was ranked 8th for reading literacy and, like the PIRLS study, the distribution of scores was wider than the international average. England was ranked 9th for mathematics, again with a wide range of scores and 4th for science with a typical range of scores (Gill, Dunn and Goddard, 2002). Children with special educational needs were excluded from the study.

When PISA was repeated in 2003, the number of schools willing to participate was not sufficient to meet the sample technical requirements and so although the results were poorer, they could not be meaningfully compared with other countries or with previous performance.

Although the PISA results from 2000 showed England well placed internationally it is worth noting that its position was indistinguishable statistically from that of Northern Ireland which was not subjected to the series of policy initiatives experienced by English schools.

TIMSS

TIMSS is conducted by the International Association for the Evaluation of Educational Achievement (IEA). It assesses the mathematics and science achievement of pupils in Grades 4 (English Year 5) and 8 (English Year 9) and is carried out on a four-year cycle.

In the latest tests conducted in 2003, for Grade 4 science just two countries performed at a significantly higher level than England, three countries performed at a similar level and 19 countries at a significantly lower level. For Grade 4 mathematics, six countries performed at a significantly higher level than England, three performed at a similar level and 14 at a significantly lower level. For Grade 8 science, 4 countries performed at a significantly higher level than England, 4 at a similar level and 37 at a significantly lower level. For Grade 8 mathematics, 9 countries performed at a significantly higher level than England, 12 at a similar level and 24 at a significantly lower level. The Grade 8 results should be viewed with extreme caution because the sample of schools did not meet the technical requirements although the number of pupils did.

England's performance over time has been monitored and compared against a group of other countries that have participated in all of the TIMSS surveys since 1995. These countries are Australia, Hong Kong, Hungary, Japan, New Zealand, Singapore, USA, Belgium (Flemish), Italy, The Netherlands and Scotland.

In Grade 4, both science and mathematics scores increased significantly from 1995 to 2003. For mathematics, the rise in mean score was larger than the average change for the comparison group. The Grade 4 pupils assessed in 2003 will have followed the Numeracy Strategy. England's mean science score also increased in comparison to the other countries in the group however this was a smaller rise but from a higher baseline position. Improvements in performance were seen across the whole range of ability.

In Grade 8, after adjusting the sample to make it as representative as possible, standards did not rise significantly in either mathematics or science.

Summary

The large-scale international studies described above enable comparisons to be made between pupils at particular points in time but it is not possible to look at the survey data in the light of information about the levels of children at an earlier point of time because to date the international studies have not incorporated a baseline assessment. In other words the international studies do not look at progress. PIRLS included a questionnaire about the readiness and awareness of print of children starting school but Clark (2004) showed problems with the approach and it is not an adequate replacement for a reliable baseline assessment. Without that common baseline it is surely hard to interpret the data generated.

In relation to changes in the attainments of pupils and the impact of government initiatives the best that can be said is:

- The PIRLS study cannot be used to comment on the impact of policies. There were too many problems with the research.
- The TIMSS study suggests a rise in maths in sciences between 1995 and 2003 in primary schools but no change in secondary schools.
- The PISA study in 2000 showed the reading, mathematical and scientific literacy of year 11 students in a positive light but could not be used to demonstrate the impact of policies.

Efforts to improve attainments

The efforts of the government to improve attainment in primary schools have been extensive. Three major reforms were introduced by the Conservative government and were continued when the Labour party came to office. The were:

The National Curriculum

Ofsted : The Office for Standards in Education

Statutory Testing of children at the end of each key stage

Additional initiatives

The list of changes from Professor Stephen Gorard, York University, in his book *Education and Social Justice: the changing composition of schools and its implications of 2000* refers to a "plethora of remedies" and "no fewer than 630 separate approaches to raising standards at Key Stage 2". An extract from his book reads:

"*In Britain there are at presently many active projects, programmes and policies intended to improve the education system. Most are in the form of initiatives to raise examination/test scores, and even where they are attempts to improve equity and justice they mostly plan to achieve this by relative improvements in examination scores. Some of these plans have appeared in the chapters of this book, including calls for the reduction of selection in GM schools, literacy and numeracy hours, citizenship, National Targets, homework clubs, male primary teachers, summer schools, teacher 'Oscars', the University for Industry, educational action zones, the abolition of the Assisted Places Scheme, individual learning accounts, a ban on calculators, and the 'national curriculum for teacher training'. A recent scheme is a revision of the GCSE examination timetable, allowing subjects with large numbers of entries to be tested later, so allowing more revision time for most candidates (Cassidy 1999h). A survey by the NfER of 245 schools found that they had introduced no fewer than 630 separate approaches to raising scores at key stage 2 alone (Sharp 1999). There are truly a plethora of remedies for education and social justice*."

Since then there have, of course, been many more changes. For example, the national literacy and national numeracy projects have been amalgamated; initiatives have been introduced to enhance attainment at KS3; a major pre-school program - Sure Start – has been established and the work of Ofsted has been cut back.

The latter is of particular interest since Ofsted was for a long time regarded as a key feature of the many levers put in place to ratchet up standards. Over many years academics and others, led by Prof Fitz-Gibbon had attacked the basis of the inspection system (the reliability and validity of judgements and its impact). Finally a paper from Newcastle University in 2003 by Shaw, Newton, Aitkin and Darnell tipped the balance. They showed that for most schools in England "OSTED inspection had no positive effect on examination achievement … if anything they made it worse".

Conclusions

Changes in standards

The general conclusions that can be drawn are as follows:

- In primary schools there have been small but clear gains during the last decade. These gains were overstated in the official statistics, but the gains are real nonetheless, although it is possible that the children's greater familiarity with the testing procedure and test formats could explain all or some of the rise. The gains have been greatest in mathematics but modest in reading. The gains from 1997 onwards continued a trend started in 1995 and steadied off in 2000 with very little advance since then although there has been some additional growth for maths.
- At the end of compulsory secondary school, at GCSE, there have also been clear gains since the Labour Party came to office. They have been steady and modest but gains nonetheless although the gains did not suddenly start in 1997.
- For both the end of primary schooling and the end of compulsory secondary education the gain in attainment over nearly a decade is small. In technical terms they both amount to an Effect Size of about 0.2. In the educational research literature many tightly controlled intervention studies have been shown to have much greater Effect Sizes and an Effect Size of 0.2 is not regarded as of great moment.
- At A level the clear gain is in the number of people who have gained A levels and moved on to higher education. But there has been a problem with standard-setting at A level. The top grades no longer mean the same as they used to. There is a suspicion that the very highest levels of attainment have been sacrificed to the greater numbers of students. The trends in numbers and leniency have been apparent for the last 15 years and did not suddenly appear after the Labour party came to office.

Impact of Initiatives: Evaluation and monitoring

The gains have been modest but the efforts have been massive. Hundreds of millions of pounds spread across hundreds of initiatives have been invested. One has to ask if the money could have been better spent. It is our opinion that many changes that were put into place without sufficient evidence of their effectiveness before they were released into schools. To take one example, when the National Literacy Strategy was being created, an evaluation with comparison groups and experimental schools was in place but before the results could be collected the urgency to launch the initiative was such that the National Literacy Strategy was pressed on schools before the evaluation could be completed and before lessons could be learned. This was very unfortunate. There can be no substitute for very careful investigation of embryonic policies before they become policy.

There is a growing movement towards evidence-based education around the world. The recent law (2001) passed through Congress in the United States, the "No Child Left Behind" Act, is probably the most influential educational legislation for a generation. It insists that that policy and action and money should be based on

initiatives that have an established evidence-base. Of the hundreds of initiatives put in by the British Government very, very few, could be said to have a strong evidence-base and in no case to our knowledge has their impact *in situ* been rigorously evaluated. One example is provided by Sure Start; a major initiative in the early years. There is good evidence that initiatives in the early years can have long-term positive impacts but we also know that the impact varies from project to project. Further most of the evidence comes from the USA. And yet Sure Start has been put in place without good evaluation mechanisms to monitor its impact. We are not well placed to learn systematically from the experiences of the Sure Start initiative.

Too often where evaluations are commissioned they are *post hoc* and lack rigour. Projects are put in place and evaluated almost as an afterthought. That is not good enough. Well-evaluated pilot studies are needed first and then on-going evaluations need to be put in place as the projects are rolled out. Lessons need to be learned initially and then continuously during the lifetime of the projects so adjustments can be made as time goes on.

Ofsted was left in place at enormous cost. It was having no positive impact and probably caused damage. It was finally changed dramatically following the paper from Newcastle University – an academic independent piece of work.

The National Literacy Strategy was put in place and then a *post hoc* evaluation costing £1M from Canada was put in place. It was too late, too expensive and too friendly.

In order to run an efficient education policy we need initiatives that are trialled rigorously using scientific methodology and then made available out to the rest of the country. Further, these trials need to demonstrate that if schools are asked to take on initiatives then we can expect improvement. Almost no initiative put in has had that kind of evidence-base to it. It has been at a higher theoretical or common sense level. Commonsense and theory do not provide a sufficient basis for national initiatives. Education is very complex, unexpected outcomes are commonplace and there is no substitute for empirical evidence. A further consideration when evaluating the impact of an intervention is the size of the effect it makes given the cost of its implementation. This can be defined that as 'Economic Significance'. A low cost intervention which results in a small positive effect might actually be more cost effective than an expensive intervention that makes a larger difference and that it is important to consider the two factors together.

A great paper by the late Donald Campbell was entitled "Reforms As Experiments". It suggests that politicians need to be helped out of the trap in which they find themselves – they initiate reforms and are then held accountable for the results. This is not a good system on which to base the nation's education. We need to find a way forward in which politics does not get involved in the minutiae of educational decision-making and in which reforms can be introduced on a shared understanding that changes, even U-turns, are expected. Above all we need to depoliticise education and to ensure that national initiatives are thoroughly trialled and continuously independently monitored when they are introduced.

References

CEM Centre website: http://www.cemcentre.org/news/standards.asp

Davies, J. and Brember, I. (1997) "Monitoring Reading Standards in Year 6: a seven year cross-sectional study", *British Educational Research Journal*, **23**(5): 615-622.

Davies, J. and Brember, I. (1999) "Standards in Mathematics in Years 2 and 6: a nine year cross-sectional study". *Educational Review* **51**(3): 243.

Davies, J. and Brember, I. (2001) A decade of change: Monitoring reading and mathematics attainment in Year 6 over the first ten years of the Education Reform Act, *Research in Education*, 65 (May) pp 31-40.

Gill, B., Dunn. M. and Goddard, E. (2002) *Student Achievement in England: Results in Reading, Mathematical and Scientific Literacy Among 15-Year-Olds From OECD PISA 2000 Study.* London: The Stationary Office.

Gorard, S. (2000) *Education and Social Justice: the changing composition of schools and its implications*. Cardiff: University of Wales Press.

Clark, M.M. (2004) International Studies of Reading, Such as PIRLS – A Cautionary Tale. *Education Journal*, 75, p25 – 27.

Hilton, M. (2001) *"Are the Key Stage Two Reading Tests becoming easier each year?"* Reading (April 2001), pp 4-11.

Hilton, M. (Submitted for publication) *The Measuring of Standards in Primary English since 1996, the PIRLS Report and the NC Test: Issues of Validity and Accountability*.

Labour Party (2003) http://www.labour.org.uk/news/primaryschools

Massey, A., Green, S., Dexter, T. and 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.

Minnis, M. and Higgs, S. (2001) *Evaluation of the National Literacy and Numeracy Strategies: Technical report for the Testing Programme 1999- 2001*, Downloaded from the QCA website.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. and Kennedy, A.M. (2001) *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary School in 35 Countries.* Boston MA, International Study Center, Lynch School of Education, Boston College.

PISA (2003) www.pisa.oecd.org

Statistics Commission (2005) *Measuring Standards in English Primary Schools: Report by the Statistics Commission on an article by Peter Tymms*. pp. 9. London: Statistics Commission.

TIMSS (2003) www.timss.bc.edu

Twist, L., Sainsbury, M., Woodthorpe, A. and Whetton, C. (2003) *Reading All Over The World: PIRLS Report For England.* Slough, National Foundation For Educational Research and London, DfES.

Tymms, P. and Fitz-Gibbon, C.T. (2001) Standards, Achievement and Educational Performance, 1976-2001: A Cause for Celebration? *Education, Reform and the State: Politics, Policy and Practice, 1976-2001*. R. Phillips and J. Furlong. London, Routledge.

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

Tymms, P., Merrell, C. and Jones, P. (2004) Using Baseline Assessment Data To Make International Comparisons, *British Educational Research Journal*, **30**(5) p673 – 689.