Curriculum, Evaluation and Management Centre

STANDARDS, ACHIEVEMENT AND EDUCATIONAL PERFORMANCE: A CAUSE FOR CELEBRATION?

Peter Tymms
and
Carol Fitz-Gibbon

In Robert Phillips and John Furlong (Eds.)
Education, Reform and the State:
Twenty-five years of politics, policy and practice
(2001) London: RoutledgeFalmer.
Chapter 11, pp. 156-173

11 Standards, achievement and educational performance
A cause for celebration?

Peter Tymms and Carol Fitz-Gibbon

Introduction
This chapter considers one of the most important and controversial issues relating to the period 1976-2001, namely the question of educational standards in primary and secondary schools in England and Wales. It begins by exploring the inter-relationship between educational reform and the drive for higher educational standards. Using a number of sources of data, it analyses standards in both primary and secondary schools during the period. The chapter not only attempts to make some tentative judgements about whether educational standards can be said to have risen or fallen during the period, but also raises issues about the ways in which performance is measured in England and Wales.

Politics, standards and the imperative of reform
As the other chapters in this volume amply demonstrate, the last twenty-five years have seen remarkable changes in education in England and Wales. The period has seen the introduction of a statutory curriculum, as well as fundamental changes to the ways in which schools are organised, managed and controlled. As Broadfoot has shown in her chapter, the last fifteen or so of those years have witnessed increasing attempts to monitor and control state schools through statutory assessments. The major reason for the statutory changes was a political supposition – initiated first by the authors of the Black Papers and continued by Callaghan at Ruskin in 1976 – that all was not well in schools and that, in particular, educational standards were low and falling.

It could be argued, therefore, that the essential driving force for many of the reforms during this period was putatively a drive for standards (Shorrocks-Taylor, 1999). A fairly simple hypothesis guided the drive for reform: economic success would be enhanced by increasing the levels of skills and qualifications that are needed to compete in a changing world; secondary schools should deliver these skills and qualifications; pupils in primary schools would have to reach certain levels in the basics to benefit from secondary education. Therefore, priority had to be given to the basics in primary schools (Blunkett, 1997).

Until just over a decade ago, the main external assessments were Advanced level (A level) examinations, designed originally to select pupils for university, and Ordinary level (O level), to which the Certificate of Secondary Education (CSE) had been added for the less able pupils. However, the fact that there was a dual system for pupils at the age of 16 was increasingly seen as a problem, and from 1988 onwards the CSE and O-level examinations were replaced by a single system, the General Certificate of Secondary Education (GCSE) in 1988. Both A levels and GCSEs were 'high stakes' in the sense of controlling access to employment and further and higher education, yet as a means of monitoring standards in particular schools and in the nation as a whole they were of limited value. Moreover, although there were clear views about what constituted 'good' schools in particular areas, there were no league tables and LEAs did not monitor the results by school.

In 1987, the Conservative government therefore initiated the extension of formal assessment to all pupils of ages 7, 11, 14 and 16. In order to facilitate the change, Kenneth Baker, the then Secretary of State for Education, established the Task Group on Assessment and Testing (TGAT). Its task was to devise a series of statutory assessments that would operate at the

ends of what have now become known as Key Stages 1, 2 and 3; the final Key Stage of compulsory education – Key Stage 4 (KS4) – coincides with the end of GCSE. The Education Reform Act of 1988, which introduced the National Curriculum, ensured that these new assessments were phased in and started to work their way through the system in parallel with the new curriculum. The assessments took a while to become stabilised, as the pure criterion referenced approach proved to be unworkable and as some of the pupil tasks similarly failed to live up to expectations. But these Key Stage tests have now become part of the established education scene in England and Wales. A further addition to statutory assessment is the 'baseline assessment' of pupils on entry to school, which was introduced in 1999.

As a result of these changes, there is now in place a system of national testing of children's achievement throughout the school system, and the results are starting to generate an extensive longitudinal database. These data can now be added to findings from a series of earlier surveys of educational attainment. The National Foundation for Educational Research (NFER), for example, conducted repeated surveys of reading, and some of this work goes back to 1948 (Brooks *et al.*, 1995). The Assessment of Performance Unit (APU), which was created in 1976, also contributed to the monitoring of standards from 1977 until it was terminated in 1990, when the National Curriculum was introduced. And our own Curriculum Evaluation and Management (CEM) Centre has specialised in providing, directly to schools, measures of the relative progress of each student ('value added') along with a wide range of other indicators of the behavioural and attitudinal outcomes of schooling (www.cem.dur.ac.uk; Fitz-Gibbon, 1996; Tymms, 1999).

There is, therefore, very extensive data now available on children's educational achievement throughout their school careers. Given the importance attached by successive governments during this period to the nation's educational standards, it seems reasonable at this point to pose the question of how pupils' performance has changed. Following the very extensive programme of educational reform described throughout the rest of this volume, to what extent have educational standards risen?

As we will see, a first glance at the evidence does seem to confirm that standards have risen in recent years and, indeed, in some instances the rise has been truly impressive. So should such a rise be a cause for celebration? Given the new forms of assessment that have been introduced, can we be confident that standards are indeed rising? Assessing standards over time is never easy and, as we will demonstrate below, a more careful review of the monitoring procedures that have been established inexorably leads us to the conclusion that, despite the assessment reforms, it remains extremely difficult to measure changes in national standards with complete confidence. We therefore conclude our chapter with some suggestions for the future.

National Curriculum data in primary schools

Official data are available from 1995 onwards for the end of Key Stage 2 (KS2) assessments in mathematics, English and science. We shall concentrate on the percentage of pupils attaining a 'Level 4' or above. The data are summarised in Figure 11.1. This shows that in English the results have risen fairly steadily from 1995, when 48 per cent of children gained the same result. The rise seemed to be levelling off in 1998, but a sudden rise was observed the following year. The overall change is impressive and has largely been paralleled by the results of the two other core subjects – mathematics and science.

[Figure 11.1: End of primary, percentage achieving Level 4 and over]

The science results started in 1995 with a surprisingly high 70 per cent of children gaining a Level 4 or above, but this dropped to 62 per cent the following year. The drop brought the severity of science grading more into line with English and mathematics. One can speculate that the Schools Curriculum and Assessment Authority (SCAA), the body responsible for overseeing the tests at the time, decided that science was somewhat out of line with the other two subjects, and an effort was made to pull the three into line with one another in 1996. Careful analysis of the data (Tymms, 1996) showed that the three subjects were more or less aligned for more able pupils in 1995. In other words, there was little tendency for the more able pupils to get higher or lower grades in science, mathematics or English. But for the less able, higher levels were much easier to gain in science than in mathematics and English. Mathematics performance started at a very low level in 1995 (44 per cent), lower than either English or science, and this improved quite dramatically, up to 54 per cent in 1997, only to drop in 1998, possibly because a new mental arithmetic test had been introduced. Then, along with the two other subjects, it experienced a significant boost in 1999.

The data shown in Figure 11.1 raise many issues. Most of these will be discussed later in the chapter when comparative data from a variety of sources have been presented. However, a single point is emphasised here: there are difficulties associated with the use of percentages when reporting schools results and this is compounded when monitoring standards. The percentage of children who achieved a Level 4 or above tells us nothing about the children at the bottom end or top end of the distribution, and there has for some years been talk of a long "tail" in reading (see, for example, Brooks, 1998: 6). (The term "tail" refers to the group of individuals with low scores on a test, and comes from a consideration of distributions curves.)

How significant are these apparent achievements? In order to assess the magnitude of changes, various other sources of primary school data have to be brought together, which have been converted to a scale in which the mean is 100 and the standard deviation is 15. This allows the reading specialist to pass judgement using a familiar metric and also allows researchers to make assessments about the magnitude of the changes by comparing the changes with the results from carefully evaluated interventions (Cohen, 1977). Table 11.1 sets out the metrics.

Table 11.1: Changes from a score of 100

New reading score	*Characterisation in well-controlled experiments*
100	No change
103	Small change
106	Modest change
112	Large change

Figure 11.2 shows the results from a series of longitudinal studies for the reading scores of 11-year-olds. The chart includes data for the KS2 English assessment, which, of course, is not simply a test of reading, since it includes writing and spelling. The KS2 data are the same as those shown in Figure 11.1, but they have been converted to standardised scores by taking the 1995 results as being equivalent to 100. As expected, the KS2 result show a steady and impressive rise, amounting to a 9-point rise over four years. Table 11.1 suggests that this might be characterised as something between 'modest' and 'large', but that would be for well-

controlled research-based interventions. For a national change, 9 must be regarded as very large indeed.

[Figure 11.2: Reading trends]

By far the most comprehensive data set relating to standards over time comes from the USA and gives a valuable backdrop against which the English data can be seen. It comes from the National Assessment of Education Progress (NAEP) which aims to monitor standards 'by administering materials and replicating procedures from assessment to assessment' (Campbell *et al.*, 1996: 1). Over two decades there is little evidence of even modest changes in standards. The 1996 average reading scores are imperceptibly different from the scores on the same tests in 1975. Detailed investigations of the data do show some interesting changes for minority groups, but constancy is the dominant feature.

The research of Brooks *et al.* (1995: 1), using English data, demonstrates a similar finding: they comment that 'reading standards among 10/11- and 15/16-year-olds have changed little since 1945 apart from slight rises around 1950 and in the early 1980s'. The Assessment of Performance Unit (APU) surveys of reading in 1978, 1983 and 1988 confirm picture. There appears to have been a slight rise in 1983, but again the impression is one of stability. An NFER study between 1987 and 1995 suggested stability with a slight dip in 1991.

The ongoing study by Davies and Brember (1997, 1998) involves testing all Year 6 pupils in the same six randomly chosen schools within one LEA, year on year, using the same test materials. Their data show a fall between the start of the study in 1989 and 1994, followed by a rise for the next two years. These data show the largest fluctuations of the surveys, perhaps because they used the smallest samples. Nevertheless, the data do not jump around erratically over the years. Detailed investigations prompted Davies and Brember (1997: 620) to comment 'the considerable cost of implementing the National Curriculum and assessment arrangements has not appeared to result in either raising standards or producing an effective system for monitoring reading standards throughout England and Wales'.

The final set of independent data over three years is available from the Performance Indicators in Primary Schools (PIPS) project (Tymms, 1999). The project includes a half-hour reading test, the data are available from the same 155 schools, collected at the same time of year, every year, for three years. The samples involved more than 7,000 pupils each year. The results are remarkably stable and cover a period when the Key Stage results were steadily rising.

Assessing the changes in primary schools

We have, then, a considerable amount of information. Some of this information is contradictory and the challenge is to make sense of it and to make a judgement about whether standards have increased over the period as a whole. In particular, we need to assess whether the changes that have been seen in the national assessment results at the end of primary school over the last four years represent real changes in pupils' achievement. It might be that the changes shown in Figure 11.1 should be taken at face value, but on the basis of the other data presented so far there are a number of reasons to suppose that they do not tell the whole story. In particular, there are three specifics that suggest other interpretations.

First, the 1996 drop in the percentage of pupils gaining a Level 4 or above in science could be perceived to be an attempt to bring the severity of science grading into line with mathematics

and English. Second, the small drop in mathematics between 1997 and 1998 could perhaps also be explained by the introduction of a mental arithmetic test. Third, it seems reasonable to suppose that the schools were becoming used to the new testing system and beginning to coach their pupils to prepare them for the end-of-Key-Stage primary tests. This might be called 'teaching to the test'. If schools were working independently on the curriculum and were not looking at the tests, then the changes, if they can be accepted at face value, would represent a remarkable rise. If, on the other hand, the teachers were very concerned about the results and were teaching exam techniques, then what we might see is a capacity to pass tests rather than an indication of pupils being better at English, science or mathematics. This third point highlights the problems associated with the accountability/assessment system currently in existence in England and Wales.

The changes seen between 1995 and 1999 are so dramatic and so out of step with the other longitudinal data as to raise questions about their being true representations of changes in standards. The cynical view is that the Qualifications and Curriculum Authority (QCA, formerly SCAA) are quietly adjusting the cut-off marks in order to make sure that the government targets are met. We are not at all inclined to take that position; it does seem as though QCA attempt to do an honest job. We can say this with some confidence having worked with SCAA/QCA and also having read Rose *et al.*(1999).

The present English and Welsh systems of monitoring of standards over time are complicated and require that new tests be produced each year. If this were not done then some teachers would inevitably teach the items known to be on the test. In the kind of official accountability system being run for primary schools, it is not advisable to use the same tests year on year. When a new test is produced, every effort must be made to ensure equivalent standards are maintained.

How is this done? The English and Welsh approach is well documented (Quinlan and Scharaschkin, 1999), with four procedures being used. First, the same anchor test is employed each year against which the results of the new test are checked. Second, a sample of pupils who are about to take, or have just taken, the real test from one year are given next year's test (the 'second pre-testing') so that equivalent scores can be estimated. Third, expert judgements are made of the scripts. Fourth, expert judgements are made of the questions by employing the Angoff procedure (Angoff, 1971).

Judgements about the cut-off scores, which are used to define the levels, are apparently made using all four sets of information. A retrospective account of the way that the process worked in 1999 is given in Rose *et al.* (1999), which raises issues about how marks are assigned, particularly at the borderlines of different levels. Despite these concerns, we can expect the system to be quite good at ensuring comparable standards from one year to the next. Nevertheless, the consistency could never be more accurate than one mark (about 1 to 2 percentage points). On the evidence of the 1999 KS2 decision-making meeting, the variation from year to year might be as much as five marks (between 5 and 10 percentage points). This seems quite high, but it could be argued that an error rate of a few percentage points from year to year is not unreasonable.

However, in a system that always looks at the previous year to standardise the present year's data, there is a problem. There will be a drift in standards, as over several years differences will accumulate. The only guard against this drift is the anchor test, but this may eventually be thought to lose its relevance as curricula evolve and language changes. In any case, it

seems that the anchor test is not used to check standards across the years, as one might expect. It is only used to check standards from one year to the next (Quinlan and Scharaschkin, 1999:10).

A random drift may be acceptable, but the real concern is that the drift is not random. There might be a tendency for borderline decisions to be influenced by targets – especially with the data on the consequences of their decisions being explicitly made available at decision-making meetings. There is also a mechanism in the present arrangements that might create a systematic movement towards lower standards (more pupils getting higher levels for the same performance). Rose *et al.* (1999) quite rightly note that the 'second pre-testing', the one on which the marks are based, is 'adrenaline-free' as the pupils know that the test is not the real thing and will not count for anything. This may result in pupils gaining lower marks on the pre-test than they would if the test were 'live'. The same lowering of standards will result from a pre-test taking earlier than the live test – pupils are younger, by the odd month, and less prepared. So when equivalents are created, standards will be lowered. This will apply not just to the statistical procedures employed but also to the script examinations. The effect does not have to be great for an important slipping in standards to be observed – it might simply be the odd mark every year.

Another way in which the standards procedure may not provide the nation with valid information on trends over time arises from the way the marking is carried out. The markers know that primary schools are under increasing pressure to gain Level 4s, and this might have some impact on the results. It seems unlikely that this would have a large effect, since markers might be expected to try to mark as accurately as they can, but it would result in a steady upward trend. Of more significance is the effect of the new proposal of returning scripts to schools. This is an innovative and positive step, which is to be welcome. However, over the years markers are increasingly likely to find that their judgements are challenged. The challenges are unlikely to be for errors that result in higher marks, and this may mean that an additional mark here and there might be given where before it was not.

We have, then, four mechanisms by which standards at KS2 might have been lowered over the years:

- Decision-making about cut-off marks may be influenced by knowledge of the impact that they will have on published 'high stakes' targets.
- Pre-test results are equated to live results, and yet the pre-test is adrenaline-free and taken earlier than the live test.
- Markers may be looking more kindly on borderline cases as the stakes are raised.
- The return of scripts to schools may increasingly involve challenges and a reaction among markers to give pupils the benefit of the doubt.

If any of these mechanisms are in operation, then one should ask why it is that similar drifting has not been noticed at Key Stage 3 (KS3). The simple answer is that there has been a fairly steady, but slow, drift upwards in mathematics and English, although not in science. Further, the pressure at KS3 has not been as acute as at KS2, and all four mechanisms might not be expected to operate with such bite. Finally, the 'second pre-testing' may not have followed the same timing for the two key stages. It is worth noting at this point that the large gains in 1996 in English and mathematics (Figure 11.1) at KS2 were followed by a very slight drop and a very slight rise respectively by the same cohorts when they reached KS3 in 1999. The

simple idea that better results in primary school will be followed by secondary gains has not been borne out.

The kind of monitoring structure that has been put in place nationally, then, driven as it is by accountability, cannot guard against drifting standards. The only way to be sure that consistency is maintained is to employ the same secret questions, administered under the same conditions, to equivalent samples of pupils, on almost the same date over many years. Changes in language and in the curriculum always have to be taken into consideration. But when the data displayed in Figure 11.1 are taken into account, one certainly would not expect the English language to have changed much over five years, and the kinds of reading skills needed for children to access the secondary curriculum have hardly changed. We can reflect at this point that to have had yearly independent representative samples of pupils assessed on the same test at the same time each year would have been invaluable.

Standards over time at ages 16 and 18

For external examinations taken at the ages of 16 and 18, we can apply similar approaches. Do the data sets that are available indicate a possible change in 'standards', and if so how might this have happened? The answers again require careful scrutiny of available data.

Two major features of the data are that enrolments in advanced courses have increased and the grades have increased. In 1955 only 13 per cent continued directly into the sixth form or some type of post-16 provision, whereas twenty years later, in 1975, the figure had almost trebled to 38 per cent, and then in the next twenty years it almost doubled to 72 per cent. The proportions of students going into higher education showed similar growth in the same twenty-year periods, from 4 per cent to 31 per cent in 1955, 1975 and 1995 respectively.

During these substantial increases in the proportions of students continuing in education, the failure rate declined. At A level the failure rate was about 30 per cent in almost all subjects until 1986. This rate was recommended by the then Secondary School Examinations Council, formed by the university-led examination boards. Not only were almost one in three candidates failed in most A-level subjects, but also a further 20 per cent were given an E grade. Thus a D grade was a grade in the top half of the distribution for many years at A level (Fitz-Gibbon, 1985: 55; SCAA/Ofsted, 1996).

Latin was exempt from the recommended 30 per cent failure rate because it was recognised that the entrants to A-level Latin were an exceptionally able group; to fail 30 per cent would be to set a very high standard indeed for A-level Latin. However, differences in the ranges of ability of candidates choosing other subjects were not taken into account. This resulted in unrecognised differences in difficulty between A-level subjects. For example, in 1988 students taking A-level mathematics were on average awarded a grade that was two grades lower than those obtained by similar students who had opted for A-level English (Fitz-Gibbon, 1988). Lack of knowledge of this general national pattern led many headteachers to blame mathematics departments for having lower grades than the English department. Indeed, it was such a situation that led to the development of the A-level Information System (ALIS) in 1983 (see Fitz-Gibbon, 1996: chapter 7).

Subsequent research for SCAA/QCA showed that the sciences, mathematics and foreign languages also attracted more able students, as did Latin, but since no adjustments were made to the failure rates, unlike Latin, they were all severely graded compared with most other subjects (Fitz-Gibbon and Vincent, 1994). The research study for SCAA used four methods

to compare the difficulties of various subjects: comparing the grades achieved by individual students taking the same 'hard' and 'easy' subjects; comparing grades awarded to large samples of students who had the same previous levels of achievement; comparing grades awarded to large samples of students who had the same previous levels on a measure of developed abilities; and using comparisons with the collection of grades each student obtained in all the subjects taken, using the method of 'relative ratings' (Kelly, 1976).

It was quite clear from all four methods that A-level grades could not be considered equivalent to each other. Thus, a B grade in one subject might be obtained on average by students who in another subject would average a C grade. This was important, given the growing agenda for judging schools on the basis of examination data, and the situation helped to stimulate interest in the use of 'value added' (i.e. measures of relative progress). If examinations results were heavily determined by the difficulty levels of the subjects taken and by the intake to the school, then these factors needed to be taken into account in judging schools.

The data in the Fitz-Gibbon and Vincent (1994) report mentioned above were based on schools that voluntarily participated in the A-level Information System, and it was possible that this voluntary sample was not nationally representative. However, Sir Ron Dearing had the DfEE check the findings using 100 per cent national samples, and the findings were broadly confirmed. For example:

> Replication of the Fitz-Gibbon and Vincent work shows that, broadly speaking, the general pattern of variation across subjects reported by these authors is confirmed by the national data ….. in terms of the value-added from GCSE to 'A'-level, boys made greater progress than girls and ….. the maximum value-added scores were obtained in the arts subjects. Again, these findings are confirmed by the national data.
>
> (Dearing, 1996: 4-5)

The concept of subject 'difficulty' is intimately linked with 'standards'. More students achieving higher grades could be interpreted as the result of better teaching, more effort on the part of the students, more use of private tutors, and other such factors that might have led to 'really' improved achievement. But, on the other hand, the grade inflation could be due to easier examinations or easier grading. The former explanation would denote improving standards, and the latter would denote falling standards. The reality might be a mixture of declining difficulties and improved learning.

Are we seeing grade inflation, where an A grade is now 'worth' less than previously? Or are we seeing many more students reaching the original high standards that put the UK in very highly ranked positions in international comparisons in the sciences (Smithers and Robinson, 1991)? Have standards fallen or risen?

There is no doubt regarding the fact that more students have been achieving higher grades in recent years, and not only at A level. There have been similar grade changes at GCSE. Indeed, the introduction of an A* grade (higher than an A) in 1994 was clearly symptomatic. Table 11.2 shows the decreases in the percentage of students obtaining low grades and the increases in the percentage of students obtaining high grades between 1988 and 1995, using GCSE candidates of all ages in England, Wales and Northern Ireland. Additional data in the same report showed the percentage achieving five grades of C or better rose from 23 per cent in 1975 to 43 per cent in 1995 (SCAA, 1996: figure 6).

Table 11.2: Percentages of candidates awarded various grades: GCSE in 1995

Grade awarded	*U*	*G*	*F*	*E*	*D*	*C*	*B*	*A*[a]
For all candidates and subjects, 1995	1.7%	4.7%	9%	14%	18%	22%	18%	10%
Change in % from 1988 to 1995	-1.8	-2.1	-3.0	-2.0	0.1	2.2	5.5	5.0

Notes:
Based upon data in SCAA 1996
[a] A and A* grades combined

The data are not in doubt. The difficulty that faces us lies in the interpretation: distinguishing between declining standards and improved actual achievements. One SCAA report stated that 'No attempt is made within the report to interpret the findings' (SCAA, 1996: introduction). Another SCAA report, prepared jointly with Ofsted, investigated three subjects (English, mathematics and chemistry) at GCSE and A level. It noted that many changes had taken place over the years in syllabuses, content, question type, coursework and question structuring, making 'comparing standards over time a complex task' (SCAA/OFSTED, 1996: 10). It suggested that Codes of Practice now ensure 'greater consistency of procedures' between examination boards, but conceded:

> It is possible that the emphasis given to awarders' judgement of the quality of candidates' work rather than to statistical data, coupled with a tendency to choose the lower of two scores when there is a decision to be made about setting the minimum mark for a grade, *may have allowed small, unintended but cumulative reductions in grade standards in successive years*.
>
> (ibid.: 15, emphasis added)

This frank admission regarding the process of standard setting lends further credence to the points made about drifting standards earlier in this chapter. The most valuable parts of the SCAA/Ofsted (1996) report are probably the qualitative descriptions of the changing content of the syllabuses. But, given a concern with standards, these are not the topic of interest here. As the report notes: 'Standards of attainment refer to the *demands* of syllabuses' (p. 1, emphasis added). The appropriate statistical treatment is therefore to examine results in the framework of 'demand', this being understood as how easy or difficult it was for students to attain certain grades. Using this statistical monitoring approach, the aim is to make fair comparisons, whether these be from subject to subject or (the focus of this chapter) of the same subject over time. How can like to compared with like so that the changing demands can be interpreted?

That either prior achievement or a good measure of general academic aptitude predicts about 50 per cent of subsequent variation in results is now widely known as a result of decades of school effectiveness research (e.g. Madaus *et al.*, 1979; Fitz-Gibbon, 1985; Gray *et al.*, 1986; Nuttall *et al.*, 1989; Tymms, 1989; Goldstein and Thomas, 1996; Tymms, 1999). This fact permits comparisons to be made between the grades awarded to candidates who are similar in terms of prior achievement/aptitude. This framework in which predictions can be made has been called the '50 per cent framework' (Fitz-Gibbon, 1997) and it can be used now to

illustrate how subjects differ in difficulty, how these differences probably arose, and how these differences in difficulty can be examined over time.

Examples of the use of the 50 per cent framework

A 'demanding' course is one in which it is difficult to obtain a high grade, i.e. one in which grading is severe. Thus a hallmark of severe grading is that the average grade achieved by average students is low, whereas in easier subjects the average grade for average students is higher. These differences can be represented graphically, as in Figure 11.3. There we see differences between three A-level subjects, using actual data from students in 1999.

[Figure 11.3: Illustrating subject difficulties (data from 1999): business studies was easier than physics]

Each sloping line shows, roughly, the average A-level grade obtained by students with the average GCSE score shown on the horizontal axis. The business studies line is to the left because it was a subject chosen largely by lower achieving students. For business studies, the average A-level grade achieved was C for the hundreds of students who had had an average of B at age 16 (average GCSE score). In contrast, for students who had had an average of B at GCSE and took physics, we can red from the trend lines that they obtained, on average, only slightly above a D grade. Their most likely grade was thus a D. Economics was intermediate between the demanding subject, physics, and the easier subject, business studies.

If subject difficulties (or severity of grading) are not recognised, then schools will find that the lower grades in A level in mathematics, science and foreign language departments are blamed on poor teaching, whereas much, if not all, of the difference will be the result of severe grading. The 50 per cent framework and regression *segments* (rather than regression *lines*) provide the graphical representation that best captures both the range of the intake to a subject or syllabus and the difficulty or severity of grading. A regression segment is simply a trend line through the actual data. Instead of continuing the line across the whole graph, it is only as long as is appropriate to the intake.

There is a national problem with regard to monitoring standards over time using prior achievement measures (e.g. the average GCSE scores). These measures change their meaning each year. GCSE itself might be getting harder or easier. *In order to monitor standards we need a benchmark that does not alter*. We have such benchmarks, or baseline tests, in the Curriculum, Evaluation and Management (CEM) Centre. For A levels we have given the International Test of Developed Abilities (ITDA) since 1988. Data already published showed mathematics grades becoming easier to obtain from 1988 to 1994 (Fitz-Gibbon and Vincent, 1997). In Figure 11.4, that trend is shown to be continuing recently, at least in applied mechanics. Students scoring 60 per cent, for example, on the ITDA scored on average between a C and a D in 1996, but in 1999 they scored on average almost a grade higher. Other data not illustrated here show less change in English, and varying amount of change over time in other subjects, but for none of them is the trend towards the subject becoming more difficult from year to year.

[Figure 11.4: Comparing difficulties of mathematics (applied mechanics) over four years: it was easier in 1999 than in 1996]

Another reason for leaning towards a 'declining standards' interpretation at A level, rather than improved performance, is corroborating evidence from university lecturers. For those

subjects in which university admissions officers expect actual mastery of basic content, such as mathematics or chemistry, lecturers often have views as to the extent to which students with high grades do or do not know as much as in previous years. This source of professional judgement may be unreliable but we would be inclined to believe otherwise, particularly in those cases where thee is corroborating evidence. For example, the Royal Society of Chemistry published extensive records of the same test given across may years (Barber *et al.*, 1994). Furthermore, opinion among mathematics lecturers is quite consistent, and there is a rather measurable subject.

Turning to GCSE results over several years, we also find many regression segments floating up the page (i.e. indicating easier grading) for some subjects when plotted against an unchanging baseline. This suggests that the GCSE examinations in some subjects were becoming easier. In no subjects do they appear to have been becoming more difficult. The suspicion of falling standards is aroused, though not proved.

The changes at A level and a GCSE have not been uniform across all subjects. The difficulty or demand of English has changed very little, whereas there have been larger changes in mathematics and the physical sciences, subjects that were severely graded for many years. They are now more in line with other subject. Could it be that, despite the upwardly floating regression segments, thee has in fact been no change in standards, just a great leap forward in student achievement? Could it be that the target-setting agenda, the inspection system and the publishing of examination results in 'league tables' has led to much hard work, and some of the 'fall in standards' is actually higher achievement? Perhaps students are working harder. If they are, this is strange, because hundreds of thousands of independent reports from students, made on confidential questionnaires completed in more than nine hundred school and colleges per year, sealed in plastic envelopes and sent to the CEM Centre, state that they are not. The amount of homework reported by A-level candidates has fallen in the very years when enrolments were increasing and grades were getting better. Although not important in primary schools, in secondary schools more homework *is* associated with higher achievement. It would therefore be surprising if there were real improvements in achievement when the amount of homework reported is declining.

Judgements about grade inflation

It seems clear in the external examinations at age 16 and 18 that there has been a less severe grading, particularly in mathematics and the sciences. This is not necessarily regrettable at all – indeed, an examination system that produced large numbers of failing grades after honest effort by teachers and students would be wrong; that used to be the case at A levels, and objections were raised at the time (Fitz-Gibbon, 1985).

Furthermore, a policy of inclusion demands the opportunity to obtain qualifications that are suitable for a wide range of aptitudes. It should be noted that the inclusion of lower socio-economic status students does not of itself require much change in standards since, rather than the 50 per cent prediction from prior achievement/aptitude, socio-economic status (SES) measures predict only 9 per cent of the variation. The change in standards is driven by the increased staying on of less able students, only some of whom are from low SES backgrounds. Many from such backgrounds are very able.

There is another policy reason for reducing the difficulty of mathematics, science and foreign language subjects. These subjects are seen as contributing to our economic competitiveness: were they allowed to remain much more difficult than other subjects, this would drive

schools towards steering students away from such subjects, as it might be feared that the lower grades could damage their standing in the School Performance Tables. However, the situation is more subtle, since high grades may be more reliably obtained in the more quantitative subjects.

The inflation cannot go on indefinitely. Already an extra grade has been added at the top end of the GCSEs, an A*, better than an A. Where is it to end, and will every Secretary of State stake his or her future on ever-increasing grades, as the current one has done? Some non-verbal aptitudes of students have slowly increased by as much as an Effect Size of 1.0 in a thirty-year generation; although this effect is welcome it is not well understood. It cannot, however, account for recent rapid increases in grades, even though it might contribute some effects.

Reasonable conclusions for secondary education are that standards in external examination towards the end of secondary schooling have been adjusted downwards to meet the needs of a larger cohort and amore inclusive system. This situation reinforces the need for detailed monitoring, syllabus by syllabus, if employers and admissions officers are to interpret the data accurately and if comparisons that are made between departments in schools and between schools are to be fair.

Conclusion

We believe that many of the changes that have been introduced into schools in England and Wales over the past twenty-five years have been beneficial. Local Management of Schools (LMS), for example, involving a greater degree of autonomy at individual school level, has in many ways been an empowering development for the teaching force, a change no one seems keen to reverse. Even the development of examinations for primary schools has been welcomed by some teachers and has led to an increased amount of data for measuring performance.

But, as we have tried to demonstrate in this chapter, a serious problem surrounds the monitoring of standards over time. There is good evidence to suggest that standards have not been maintained at A level – and yet the extent of the change is not known with accuracy. In primary schools there appears to have been a rise in numeracy and literacy, but it is difficult to know whether what we are seeing in the data is real or imaginary. The published results show a steady rapid rise. Further, the ways in which standards are maintained contain within them mechanisms whereby an upward drift can be expected even if standards remain steady.

Given that the period under investigation began with a call for higher standards, we find this a rather ironic state of affairs. A more accurate and reliable system of measurement is needed, as called for by a number of scholars. Davies and Brember (1997: 621) thus suggest 'a continuous policy of research is needed to try to unravel the effects of the National Curriculum and major policy changes on schools and on children's reading attainment'. The Assistant Chief Executive of SCAA responsible for testing wrote in 1996 that 'an independent benchmark could be useful in showing that standards have not slipped, *particularly if national performance improves over the years*' (Hawker, 1996, emphasis added). Foxman *et al.* (1992: 4) wrote that a system to monitor standards effectively should 'include nationally representative samples of pupils tackling appropriate tasks repeated at regular intervals ….. monitoring surveys and National Curriculum assessment could complement each other'.

We would suggest that a small group should be established with the explicit and sole task of tracking educational standards in England and Wales. This need not be an expensive group, but it needs to be – and to be seen to be – independent from government. The group will need a brief that defines its task in general terms but leaves the specifics to be worked out, since they need considerable thought and care. The terms would dictate the areas of interest (literacy, numeracy, etc.) and the ages to be studied. The dates for reports would be specified, as would the destinations for a number of the reports that would automatically be sent to a specified number of bodies, including non-governmental organisations.

England and Wales probably now have the most monitored educational systems in the world. We can rightly speak of the lessons that have been learned and of the progress that has been made. But a new structure is needed that can provide the national with crucial high-quality data on standards over time.

Acknowledgements
We wish to thank all the staff at the CEM Centre for their dedicated work, and, in particular, Dr Paul Skinner of the ALIS (A-level Information System) project and Neil Defty of the YELSIS (Years of Late Secondary Information System) project who have accepted requests for data on top of their already demanding responsibilities.

References

Angoff, W. (1971) 'Scales norms and equivalent scores', in R. Thorndike (Ed.) *Educational Measurement*, Washington DC: American Council on Education.

Barber, N., Brockington, J., and Jones, D. (1994) *Research in Assessment XI: A Skills Test Survey of Chemistry Degree Course Entrants*, London: Royal Society of Chemistry, Education Division.

Blunkett, D. (1997) Speech to the North of England Education Conference, 4 January, the Octagon Centre, Sheffield University.

Brooks, G. (1988) 'Trends in standards of literacy in the United Kingdom, 1948-1996', *Topic*, 19: 1-10.

Brooks, G., Foxman, D., and Gorman, T. (1995) *Standards in Literacy and Numeracy 1948-1994*, NCE Briefing, New Series 7.

Campbell, J.R., Voelkl, K.E., and Donahue, P.L. (1996) *Report in Brief: NAEP 1996 Trends in Academic Progress*, National Centre for Educational Statistics, http://nces.ed.gov/nationsreportcard/96report/97986.shtml.

Cohen, J. (1977) *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press.

Davies, J. and Brember, I. (1997) 'Monitoring reading standards in Year 6: a seven year cross-sectional study', *British Educational Research Journal*, 23(5): 615-22.

Davies, J. and Brember, I. (1998) 'Reading and mathematics attainments and self-esteem in years 2 and 5: an eight-year cross-sectional study', paper presented at the ECER conference, Ljubljana, September.

Dearing, R. (1996) *Review of 16-19 Qualifications*, London: SCAA.

Fitz-Gibbon, C.T. (1985) 'A-level results in comprehensive schools: the Combse projects, year 1', *Oxford Review of Education*, 11(1): 43-58 (Combse: Confidential, Measurement-based Self-Evaluation, the original name of the project later renamed ALIS, the A-level Information System).

Fitz-Gibbon, C.T. (1988) 'Recalculating the standard', *Times Educational Supplement*, 26 August: 15.

Fitz-Gibbon, C.T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness*, London: Cassell.
Fitz-Gibbon, C.T. (1997) 'Listening to students and the 50 per cent framework', in A.D. Edwards, C.T. Fitz-Gibbon, F. Hardman, R. Haywood and N. Meagher (Eds.) *Separate but Equal? A Levels and GNVQs*, London: Routledge.
Fitz-Gibbon, C.T. and Vincent, L. (1994) *Candidates' Performance in Public Examinations in Mathematics and Science*, London: SCAA.
Fitz-Gibbon, C.T. and Vincent, L. (1997) 'Difficulties regarding subject difficulties', *Oxford Review of Education*, 23(3): 291-8.
Foxman, D., Gorman, T. and Brooks, G. (1992) *Standards in Literacy and Numeracy*, NCE Briefing 10.
Goldstein, H. and Thomas, S. (1996) 'Using examination results as indicators of school and college performance', *Journal of the Royal Statistical Society*, 159(1): 149-65.
Gray, J., Jesson, D. and Jones, B. (1986) 'The search for a fairer way of comparing schools' examination results', *Research Papers in Education*, 1(2): 91-122.
Hawker, D. (1996) *Can we really trust the tests? Times Educational Supplement*, 16 February.
Kelly, A. (1976) 'A study of the comparability of external examinations in different subjects', *Research in Education*, 16: 50-63.
Madaus, G.F., Kellaghan, T., Rakow, E.A. and King, D.J. (1979) 'The sensitivity of measures of school effectiveness', *Harvard Educational Review*, 49(2): 207-30.
Nuttall, D.L., Goldstein, H., Prosser, R. and Rasbash, J. (1989) 'Differential school effectiveness international', *Journal of Educational Research*, 13: 769-76.
Quinlan, M. and Scharaschkin, A. (1999) 'National Curriculum testing: problems and practicalities', paper presented at BERA Annual Conference, University of Sussex, September.
Rose, J., Downes, P., Grant, M., O'Leary, J. and Wallace, J. (1999) *Weighing the Baby, the Report of the Independent Scrutiny Panel on the Key Stage 2 National Curriculum Tests in English and Mathematics*, DfEE: http://www.dfee.gov.uk/panel/report.htm.
SCAA (1996) *GCE Results Analysis: An Analysis of the 1995 GCE Results and Trends over Time*, London: SCAA.
SCAA/Ofsted (1996) *Standards in Public Examinations 1975-1995*, London: SCAA.
Shorrocks-Taylor, D. (1999) *National Testing: Past, Present and Future*, Leicester: British Psychological Society.
Smithers, A. and Robinson, P. (1991) *Beyond Compulsory Schooling: A Numerical Picture*, London: Council of Industry and Higher Education.
Tymms, P. (1993) 'Accountability – can it be fair?' *Oxford Review of Education*, 19(3): 291-9.
Tymms, P. (1996) *The Value Added National Project Secondary Primary Technical Report: An Analysis of the 1991 Key Stage 1 Data Linked to the 1995 KS2 Data Provided by Avon LEA* (Ref: COM/96/554), London: SCAA.
Tymms, P. (1999) *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-Added Indicators*, London: David Fulton.