

Can Testing Really Raise Educational Standards?

Professorial Lecture

Harry Torrance,
Institute of Education
Manchester Metropolitan University, UK¹.

Introduction

It's a strange Alice-in-Wonderland World we enter, when we enter the world of examinations and testing; exam passes and test scores have never been higher, as I will demonstrate later in this lecture, and yet the moral panic surrounding exams and educational standards continues unabated, as a glance at any recent education headline will demonstrate. Interestingly enough however current newsworthy stories have a slightly different angle to them, compared to the usual 'shock, horror, standards are falling' headlines. Suddenly the headlines are beginning to hint that we may have too much testing, rather than too little, with the system buckling under the weight of the current load. Exam Boards have so many papers to process, and results to report, that serious errors are creeping into their procedures – Edexcel is hardly out of the news these days, for faults in their question-setting, while the Scottish Qualifications Authority had that debacle over reporting thousands of wrong results last year. Meanwhile in the Key Stage tests Headteachers are under so much pressure to deliver results that some resort to cheating, while students and teachers alike complain of the exam overload brought about by Curriculum 2000 and the introduction of A/S levels. To use an engineering metaphor, it seems that we are beginning to 'test the system to destruction'. Well, that's all very well when we want to know how much force the materials in a bridge can withstand, but it hardly seems appropriate to the future building blocks of our society – our children.

So what on earth is going on here? How did we get into this state and why as a society are we so obsessed with testing and test results? Using testing to improve schooling seems compellingly simple – set your goals, measure whether or not they have been met, do something about it if they haven't. Surely testing can indeed raise educational standards? But if so, how, and with what consequences? And if not, how else should we assess and develop our children and our schools? The UK is not alone in turning to assessment to reform and control schooling. Governments around the world have become concerned about educational standards and their implications for economic competitiveness (Torrance 1995, 1997). Likewise policymakers and educationists around the world are interested in using assessment to try to raise standards; but only in the UK, or, to be more accurate, England, has the focus become so narrow, intense and, as I will argue in this lecture, counter-productive.

In this lecture I shall be looking at:

- The emergence of 'standards' as a political issue in the UK;
- Government approaches to setting and measuring standards over the last twenty years;
- The logic of using testing to raise standards;
- The impact of national testing on standards;
- The scale and costs of the enterprise;
- What is to be done? Alternative approaches to testing, and the development of formative classroom assessment.

I will have to skim over some elements of the argument pretty quickly, but it is important to document and reflect on where this impetus for testing has come from, before moving on to

¹ Delivered at the University of Sussex, 11 June 2002.

A more fully developed published version is available as 'National Assessment in England' in Kellaghan T. and Stufflebeam D. (Eds. 2003) International Handbook of Educational Evaluation, Kluwer.

look at current impact and future prospects. I shall focus mainly on the UK but bring in research evidence from elsewhere, particularly the United States, as appropriate.

The Emergence of 'Standards' as a Political Issue in the UK

Interestingly enough, things didn't used to be like this; once upon a time, and not so very long ago – up until the late 1980s – the United Kingdom had one of the most decentralised and voluntaristic educational systems in the world; indeed in many respects it would hardly qualify as a national system at all, and this is certainly one of the main reasons why the pendulum has swung so far in the opposite direction.

Following the 1944 Education Act, and up to the 1988 Act which introduced the National Curriculum and National Testing, Local Education Authorities (LEAs) were responsible for providing education in their locality. There were no generally agreed curriculum goals for the system, far less centrally mandated subjects and programmes of study; no general system of assessment in primary schools; and only GCE O-level and CSE examination syllabuses to guide teaching and assess performance in secondary schools². And even the Secondary Exam Boards were many and varied, with several different boards in England and Wales, along with national boards for Scotland and Northern Ireland, setting individual examination papers in individual subjects, to be taken by individual students choosing to sit as many or as few subjects as they wished.

This laissez faire system seems to have been a legacy of the peculiarly class-based nature of British, and especially English education. Up until the 1970s a high status, academically-oriented education had been largely confined to those children attending private fee-paying schools or who passed a selective examination, the 11+, to attend an academic 'grammar school' (about 20% of the age cohort). In turn, a minority of this minority proceeded to university entrance via the secondary school examination system. Most other pupils, attending secondary modern schools, left for work at age 15. The school leaving age was only raised to 16 in 1974. Thus, in effect, a series of ad hoc extensions of educational provision throughout the first half of the twentieth century, but especially after the second world war, created a very piecemeal educational system in the UK. An academic education was thought appropriate for a small elite likely to progress into social and economic leadership roles. A non-academic vocationally oriented education was thought appropriate for the rest. Historically then, and this is true of other countries besides Britain, assessment was used for, and was focussed upon, the selection and certification of individual pupils, rather than having any overall role in controlling or monitoring the system as a whole.

Criticisms of the lack of opportunity provided for the majority of pupils resulted in the gradual abolition of the 11+ and the creation of comprehensive secondary schools, throughout the late 1960s and 1970s, but without an accompanying systematic re-think of curricular provision, accreditation and qualifications, or overall evaluation. Curriculum development and evaluation activities were pursued to some degree by individual schools, LEAs, examination boards and indeed specially constituted government bodies such as the Schools Council which funded curriculum development projects; but there was no overarching monitoring or evaluation of the system as a whole, except for periodic individual inspections of schools by Her Majesty's Inspectors of Schools (HMIs). Furthermore, despite increasingly comprehensive administrative provision, O-level GCEs continued to be taken by around 20% of the secondary school population deemed able enough to aspire to higher education, while a parallel system of single subject CSEs was developed for the next 40% of the ability range, with only the highest grade in CSE (grade 1) being considered equivalent to the lowest pass grade in GCE O-level (grade C). Thus very little coherent policy discussion took place before the late 1970s with respect to what sort of overall curricular provision might be appropriate for a truly comprehensive system of education, far less how this might be assessed at individual level or evaluated at national level.

² GCE - General Certificate of Education which could be taken at O-level or A-level and is still taken at A-level; and CSE - the Certificate of Secondary Education - introduced in 1963 and abolished with the advent of GCSE in 1988.

In parallel with these developments the 'terms of trade' were turning decisively against Britain's old primary and secondary manufacturing industries (coal mining, steel making, shipbuilding, etc.) especially after the so-called 'oil crisis' of 1974 when prices were raised significantly. Unskilled jobs began to disappear rapidly and suddenly unemployment, and especially youth unemployment, became a major social and political concern. This rise in youth unemployment in the 1970s was coincidental with the growth of comprehensive education and falling educational standards were identified as a major part of the problem. The argument was put that young people needed to come out of the compulsory education system with higher (possibly different) levels of achievement in order to gain more skilled employment.

Except that politicians had no hard evidence about overall educational standards – high, low, or indifferent. To reiterate, O-level and CSE were run by many different Exam Boards raising issues of comparability, and were not taken by many pupils anyway. The 11+ was being phased out as the key test at the end of primary schooling, and again, in any case, it was by no means a universally applied test; rather it was devised and administered locally by LEAs and pass rates varied enormously around the country, in direct proportion to number of grammar schools places available – one of the reasons why it fell into disrepute in the first place (Torrance 1981).

Thus began the so-called 'Great Debate' of the mid 1970s, launched by a speech by the then Prime Minister James Callaghan in 1976, and it's a debate we've been having pretty much continually ever since; a debate about setting national standards for educational provision and achievement, and how to measure them.

Government Approaches to Setting and Measuring Standards

So the issue, in an increasingly comprehensivised system aspiring to underpin national economic development rather than simply select a social elite, was how to establish national curriculum goals and how to monitor and improve standards of achievement. Interestingly enough, from the mid 1970s through to the late 1980s, direct central government intervention was seen as neither feasible nor desirable. Government's responsibility was to stimulate debate and guide development through policy statements and setting a framework of expectation. Thus for example a government consultative document published in 1977, following Callaghan's speech, argued that:

the time has come to try to establish generally accepted principles for the composition of the...curriculum for all pupils...there is a need to investigate the part which might be played by a 'protected' or 'core' element of the curriculum common to all schools (DES 1977, p.11).

So the focus was very much on the curriculum. The document also argued that "a coherent and soundly-based means of assessment for the educational system" was necessary, but explicitly rejected the production of 'league tables' since "test results in isolation can be seriously misleading" (Ibid. p.17). Furthermore it stated that:

It has been suggested that individual pupils should at certain ages take external 'tests of basic literacy and numeracy', the implication being that those tests should be of national character and universally applied. The Secretaries of State reject this view...the temptation for schools to coach for such tests would risk distorting the curriculum and possibly lowering rather than raising average standards. (DES 1977, p.18)

Nor was this reticence (some might say, good sense) confined to previous Labour governments. Nearly ten years later, in 1986, just as GCSE was being introduced, Sir Keith Joseph, the then Secretary of State for Education in Mrs. Thatcher's Conservative government, argued that it was vital to secure "national agreement on curricular objectives" and that "meaningful assessment is possible only to the extent that there is...agreement...about objectives" (Joseph 1986, p. 181 & 182); but added:

It is easy to say...that we wish to start assessing standards in the most basic aspects of education such as reading comprehension and arithmetic computation. But [research]...has shown that apparently slight changes in the phrasing of questions can produce widely differing results...[]...we need a wide range of instruments...because assessment has many important aims and we cannot expect a single form of assessment to encompass them all equally well
(ibid. pp. 179/180)³.

In fact the process of defining and establishing national goals and objectives began in earnest under Sir Keith Joseph with the introduction of GCSE⁴ to replace O-level and CSE. In particular the laying down of national criteria and subject-specific criteria by the Government's Secondary Examinations Council, with which all of the new GCSE syllabuses had to comply, established the framework for the clear national goals which were being sought.

These criteria only applied to the latter stages of secondary schooling however, and clearly, government became impatient with the pace of change. Moreover the educational debate became encompassed within a wider debate about consumer interests versus producer interests, and this is crucial to understanding the change of emphasis in government policy. By the late 1980s teachers, and educationists more generally, were no longer seen as autonomous professionals legitimately contributing to debate and development, but as an obstructive self-interested producer group which had to be brought to heel. Discussion was no longer invited, the DES decided it knew best:

The Government now wishes to move ahead at a faster pace...[]...consistent improvement in standards can be guaranteed only with a national framework for the...curriculum...backed by law (DES 1987, pp. 3 & 5)

The argument, which, with various further embellishments with respect to target setting and publishing results still summaries the position today, was that:

A national curriculum backed by clear assessment arrangements will help to raise standards of attainment by:

- (i) ensuring that all pupils study a broad and balanced range of subjects...
- (ii) setting clear objectives for what children...should be able to achieve...
- (iii) ensuring that all pupils...have access to...the same...programmes of study which include the key content, skills and processes which they need to learn...
- (iv) checking on progress towards those objectives and performance at various stages...

(DES 1987, pp.3-4)

Interestingly enough, the DES further claimed that this was "a proven and essential way towards raising standards " (ibid. 1987 p.10); something which Sir Keith Joseph hadn't claimed only a year before, and indeed is highly contestable.

The Logic of Using Testing to Raise Standards

So what is the logic of setting and measuring standards; what sort of 'theory of school improvement' is implicitly appealed to by testing children and publishing results? Furthermore, what is the evidence, both in the UK and internationally, for using testing to raise standards; and what are the consequences?

The basic theory, in-so-far as it can be discerned is that:

³ His speech referred here to the work of the Assessment of Performance Unit, a standards monitoring unit set up by the previous Labour Government in 1974. The exact phrase reads: "But the work of the APU has shown that apparently slight changes in the phrasing of questions can produce widely differing results".

⁴ GCSE - General Certificate of Secondary Education

- i) agreement can (and should) be reached on national educational aims and specific curricular objectives;
- ii) these can (and should) be expressed in traditional subject-specific programmes of study;
- iii) in turn, tests can (and should) be devised which validly assess such objectives and accurately represent pupil attainment of them;
- iv) setting targets, publishing results and constantly pressing for increases in scores will both raise standards and provide evidence of these rising standards.

All of these steps in the process are questionable, especially the assumption that a curriculum for the 21st century can be written and imposed, once-and-for-all, in the last decade of the 20th century. It is also very problematic to try to capture all the goals of education in a subject-based curriculum. Knowledge does not have to be codified and presented in traditional subject terms; the curriculum could be organised around, for example, more integrated notions of 'problem solving', 'community study', 'scientific method', 'aesthetic experience', 'personal and social development' and so on. Certainly, many would argue that the job of schools is far wider than subject-specific teaching, and indeed at various times government policy and legislation has referred to broader goals of the spiritual, social and emotional development of pupils. Unfortunately they do not receive the high profile attention paid to subject-specific test results and as a very experienced American researcher has so succinctly put it – "you get what you assess, you don't get what you don't assess" (Resnick & Resnick 1992, p. 59). Likewise, developing valid measures of even what you do decide to assess is a "sophisticated exercise that...will be difficult...[and may] not be wholly achievable" as Sir Keith Joseph also recognised in his 1986 speech (p.183).

I'll return to these issues later. But for the present we should also note the consequences of curriculum overload which flowed from the byzantine process of National Curriculum development and implementation. I don't intend to dwell on what was in effect 10 years of constant upheaval in the school curriculum, especially the primary school curriculum, but it is now the case that many of the assumptions enshrined in original national curriculum and national testing policy have been unpicked by successive revisions such as those proposed by the Dearing Review (1993), and now the new government Green Paper suggesting more choice and flexibility post-14 (DfES 2002).

But the key issue to be addressed here is the focus on testing – the policy claim that tests can encapsulate standards, communicate them unambiguously and raise them by exerting pressure on the system.

Were standards too low?

A good starting place to get some sort of purchase on the issues at stake is to explore what data existed in the system prior to 1988. What evidence is there that standards were too low? As noted already, the short answer is 'precious little' since very little easily comparable national data existed in the system. Interestingly enough however, what data as did exist indicated that standards were rising throughout the 1970s and 1980s and have continued to rise since.

Since its introduction, CSE grade 4 had always been thought of as a reasonable level of attainment to be expected of the 'average' pupil. And when GCSE was first announced the intention of policy was:

bringing the level of attainment of at least 80 to 90 per cent of all pupils up to at least the level currently associated with the average, as reflected in CSE grade 4
(DES 1985, quoted in Kingdon & Stobart 1988; cf also DES 1986 p.43).

Now of course “CSE grade 4” means very little in terms of any absolute standard. The nature and content of the UK examination system has changed significantly over time, as outlined above. And ultimately, as Mike Cresswell, head of research at the Associated Examining Board (AEB), has noted:

Standards...are social constructs created by special groups of judges...who are empowered through the examining boards as government-regulated social institutions to evaluate the quality of students’ attainments on behalf of society as a whole.

(Cresswell, 2000, p. 5)

So there are no ‘absolute standards’ – only what experienced examiners judge to be appropriate at a particular point in time.

But insofar as we can make any sort of judgements from the official statistics, in terms of broad equivalences, we can note that successive policy reports have suggested linking the old GCE O-level ‘pass’ (grade C) to CSE grade 1 and GCSE grade C. CSE grade 4 thus links to GCSE grade F:

Figure 1

<u>GCE</u>	<u>CSE</u>	<u>GCSE</u>
		A*
A		A
B		B
C	1	C
D	2	D
E	3	E
F	4	F
G	5	G

Given the DES’s aspiration in 1985, we might imagine that large number of pupils were not achieving at least CSE grade 4 in the early 1980s, and we might look to see how many are achieving at least GCSE grade F today. In fact official statistics are usually reported in terms of pupils achieving at least 5 A*-Cs and 5 A*-Gs, so it is difficult to make exact comparisons and one would have to go back through individual subject results to calculate the percentage of A-F grades. However, if we note that in 2000, for example, only 4.8% of entries in Maths were graded ‘G’, and 2.2% in English were graded ‘G’, we can probably treat the 5 A*-G statistics as accurate to within 3 or 4 percentage points – sufficient to the purpose here.

And we can see from Table 1 that considerable progress towards the government’s target had *already* been made between 1975 and 1985. And the original target set by Sir Keith Joseph has been exceeded for some years now.

Table 1

% of pupils gaining O-level/CSE/GCSE 1974/5 – 1999/2000

	% 5 or more A*-C	% 5 or more A*-G
1974/5	22.6	58.6
1984/5	26.9	74.3
1987/8	29.9	74.7
1994/5	43.5	85.7
1999/2000	49.2	88.9

Source: DfES 2001 (results for England)

NB 1987/88 is first results of GCSE;

Grades A*-C includes O-level A-C, CSE grade 1 & GCSE A*-C;

Grades A*-G includes O-level A-E, CSE grades 1-5 & GCSE A*-G

Similarly, focussing on English and Maths results since 1980 and particularly since the introduction of GCSE in 1986 (first results 1988) we can see pass rates steadily rising (Table 2).

Table 2

% pupils gaining O-level/CSE/GCSE 'passes' (Grades A*-C) in English and Maths since 1980, including 1987, the final year of the dual system and 1988 the first year of GCSE results

	English: A*-C	Maths: A*-C
1980	32.6	31.9
1987	46.2	37.1
1988	48.9	41.9
1995	55.6	44.8
2000	58.6	50.4

Sources: DES (1980), DfEE (1996), DfES (2001a) (results for England)

NB: DfES, 2001b (SFR45/2001) confirms these trends of marginal increases for results in 2001.

Of course, some critics, when results go up, simply complain that the examination(s) have got easier. Thus, if results go down, standards are falling; and if results go up...standards are falling.

Likewise, governments can always argue that no matter how high standards are, they need to be higher still, in order to maintain international economic competitiveness – we are in a race which demands eternal vigilance and effort. And certainly, if we assume a reasonable indication of a sound secondary education is 5 GCSE A*-C grades, including English and Maths (considerably more demanding than the DES policy announcement of the 1980s), then around 50% of the school population is still not achieving this. So there are certainly arguments to be made about continually striving to raise educational standards, but these arguments are rarely put positively – standards are always said to be too low – and rarely address the actual evidence of significant progress which has been made over many years. The plain fact of the matter is that 20 years ago only 25-30% of pupils were passing the exams that 50% now pass. More young people are passing more exams than ever before in our history. And the trend started long before the introduction of the National Curriculum and National Testing. Indeed, looking back at these figures, GCSE with its more interesting syllabuses and more varied assessment methods seems to have been the real catalyst to sustained improvement. Clear curriculum goals combined with different assessment methods, particularly coursework assessment and practical assessment, and clear criteria by which to establish positive achievement, seems to have been the main reason for this rise in pass rates throughout the 1980s and 1990s.

The Impact of National Testing

Whether or not *any* increase in test scores is an unambiguously good thing, however, remains an issue, because of the backwash effect on the curriculum which, particularly, National Testing seems to have brought in.

Currently, National Curriculum Test scores are rising and primary schools are well on the way to meeting the government's targets for numbers of pupils reaching level 4 in National Curriculum English (80% target by 2002) and Maths (75% by 2002) at age 11:

Table 3

% pupils gaining National Curriculum Assessment level 2 or above at KS1 (age 7) and level 4 or above at KS2 (age 11)

	KS1		KS2	
	English	Maths	English	Maths
1992	77	78		
1995	76	78	48	44
1996	80	80	58	54
2000	81/84	90	75	72

NB 1992 is first 'full run' of KS1 tests after 'slimming down';

1995 is first full run of KS2 tests;

By 2000 KS1 English results were being reported separately in terms of attainment targets (*81% gained level 2 in Reading, 84% in Writing). Such details had been available previously but were results were routinely reported as 'whole subject' levels.

Sources: The Times, 26/1/96; The Guardian 18/11/96; DfE (1992); DfEE (2001).

Interestingly enough, although much has been made by the government about the positive impact of the National Literacy and Numeracy Strategies (NL&NS) on these scores (e.g. BBC Education Website 15/9/99), the government's own evaluation of the NL&NS reports that the greatest increase in scores at KS2 has been in *Science*, with 85% of pupils gaining level 4 in 2000, an increase over five years of 24% compared to 19% for English and Maths). The report notes that "it is difficult to know what might account for the science results" (Earl et.al. 2001, p. 19). The implication is that the National Literacy and Numeracy Strategies cannot really be said to have led directly to rising results in English and Maths and, indeed, such findings beg some fairly severe *prima facie* questions about the validity of the whole enterprise.

What *is* most likely to have led to increases in scores, is practising for the tests. All international research evidence, gathered over many years, not to mention personal experience and common sense, suggests that this is what happens when 'high stakes' tests are encountered in educational systems; i.e. when teachers and students are faced with tests which carry significant consequences for student life chances and teacher accountability, very significant time and energy will be devoted to test preparation.

Some would argue that there is no harm in this if the tests are well-designed and encourage teachers and students to focus on worthwhile educational goals. Indeed in America a whole movement to raise standards has been built around what is termed 'measurement driven instruction' (Popham 1987). Key developers in the field, Resnick and Resnick, quoted earlier, argue that:

if we put debates, discussions essays and problem-solving into the testing system children will spend time practising those activities

(Resnick and Resnick, 1992, p.59)

But of course these are the activities which GCSE assessment encompasses, not National Testing. The main thrust of UK government policy since 1988 seems to have been much more concerned to produce simple indicators of overall progress, rather than high quality assessments to underpin high quality teaching.

Other international evidence, particularly from the United States, also demonstrates that high pressure test-based accountability systems can produce rises in test scores, but that the validity and meaning of these rises are open to serious question. Thus for example in Texas, dramatic rises in scores have been reported on a recently instigated state-wide testing curriculum and programme, the Texas Assessment of Academic Skills (TAAS). These have gained significant attention because of George Bush's election to the Presidency and the possibility that something similar may influence national US policy. However the USA also has a national monitoring programme based on light sampling across states, the National

Assessment of Educational Progress (NAEP), and when Texas scores on this national test programme are analysed they show no such gains, just slight increases in line with national trends. Researchers investigating this discrepancy between TAAS and NAEP scores suggest that the explanation lies in:

Students being coached [for]...the statewide exam...and...narrowing the curriculum to improve scores...at the expense of other important skills and subjects that are not tested

(Klein et.al. 2000, p.17).

These findings indicating “score inflation and unwanted test preparation” (ibid. p. 17) have also been reported from other states claiming similar gains in state-level test scores; and indeed such findings have been routinely reported in the American literature over many years (Atkin 1979, Linn 2000).

Interestingly enough, here in England we do not have any independent check on National Test scores. Our nearest equivalent to the American NAEP, the Assessment of Performance Unit (APU), set up in 1974, was abolished in 1989 after the National Curriculum and Assessment programme was introduced. Nevertheless all the research evidence which we do have from studies of the implementation of the National Curriculum suggests similar problems are being encountered.

To begin with we have to note that the original intention of assessing *all* National Curriculum subjects has long since been dropped, so that the emphasis is now very firmly on English, Maths and Science. Clearly this will have an impact on the balance of the curriculum, especially in primary schools. Thus while research on the introduction of the National Curriculum and National Assessment has suggested that there have been some benefits, especially in terms of an increase in collegial planning and management of the primary school curriculum (KS1 & 2), many primary school teachers have reported feeling threatened by the changing emphasis from an arts-based early years curriculum to more of a science-oriented curriculum (Pollard et.al. 1994, Croll 1996). It is even more puzzling then, that this is the area of the curriculum that shows the greatest test gains; and it constitutes further evidence that the most likely source of these gains is coaching for the tests.

Further research confirms that as the national tests have become more and more restricted in scope, and allied to target setting, so has the curriculum become narrowed:

Whole class teaching and individual pupil work increased at the expense of group work...[there was] a noticeable increase in the time spent on the core subjects...[and] teachers...put time aside for revision and mock tests...

(McNess et.al.2001, pp.12-13)

Furthermore, “in the last two years of primary schooling children encountered a sharp increase in the use of routine tests and other forms of categoric assessment” and pupils became less keen on teachers looking at and commenting on their work (ibid. p. 14; cf also Osborn et. al. 2000). This latter point, if sustained, will be very counter-productive since it is precisely by looking at and commenting on pupil work that teachers can make a positive difference to pupil learning (Black and Wiliam 1998) rather than just coaching them to improve test scores.

Nor are such concerns restricted to independent researchers, still too often dismissed as part of the self-interested educational establishment. The government’s own evaluation of the National Literacy and Numeracy Strategies noted that:

The danger is that the high political profile for...national targets may skew efforts...to increase score[s] possibly narrowing the curriculum...we have seen some evidence of this happening...in many schools considerable time and energy are focussed on test preparation...[]...squeezing out other...foundation subjects, whole-school activities and field trips...

(Evaluation of NL&N Strategies; Earl et.al. 2001, p. xi)

While the most recent annual report of the Chief HMI Mike Tomlinson warns that:

In some primary schools the arts, creative and practical subjects are receiving less attention than previously. This risks an unacceptable narrowing of the curriculum...
(OfSTED 2002, Commentary p.1)

So, we are faced with the very real possibility that as test scores go up, educational standards, defined more broadly, might actually be going down, just as the DES predicted 25 years ago (DES 1977). And to reiterate, this would not be unusual or unique to the British situation. Extensive research evidence gathered over thirty years in the United States suggests that this is exactly what does happen when high stakes, but too narrow, tests are implemented in an education system. The minimal focus of the testing programme becomes the normal curriculum maximum for teaching purposes; teachers and students alike become adept at cramming and, if the consequences of failure are too severe, cheating; and politicians move the goalposts on a regular basis anyway (cf. Atkin 1979; Smith & Shepard 1989; Linn 2001).

Other Costs

This very worrying situation also comes with other costs – both human and financial. We have just seen a significant rise in exclusions from school, up 11% from 2000 to 2001, including a 19% rise in primary school exclusions, after several years of falling exclusion numbers (BBC Education Website 23/5/02). This suggests that the current system is by no means appropriate for a growing minority of pupils and that, because of pressure to deliver results, schools are increasingly intolerant of those who are, for whatever reason, disruptive. Similarly, with respect to the teaching profession, in the short term, as noted at the start of this paper, some teachers will be caught cheating and find their careers ending in the saddest of circumstances.

The longer term consequences for the profession seem to be manifest in the perennial problem of recruitment and retention. A recent survey indicated that 12% of trainee teachers drop out of training before completion, 30% of newly qualified teachers never teach and a further 18% of new recruits leave the profession within three years (TES 2/11/01, p.1). Most explanations focus on overwork, linked to the pressure to meet targets, along with relatively low pay for an all-graduate profession. But lack of control over curriculum and teaching methods is also likely to be important. There is very little scope in the system for innovative educators who might wish to experiment with new curricular programmes at local level. And, coupled with the immediate crisis of teacher numbers, one wonders where the creative leaders of the profession will come from in five or ten years time.

Another key issue is financial cost. Approximately 605,000 pupils took KS1 tests in England in 2000; 620,000 took KS2 tests; and 580,000 took KS3 tests (DfEE 2001). Multiply these figures by the number of subjects and the number of papers in each subject and the figures run into millions - all to be set, printed, distributed, collected and marked annually. At GCSE (KS4) there were 554,000 pupils in 2000, entering a total of 4,744,861 GCSE examinations (DfES 2001). Every year therefore, within the compulsory school system, around 2.3 or 2.4 million schoolchildren are sitting in excess of ten million separate national tests and examinations.

Table 4

Number of pupils taking tests in 2000

	KS1	KS2	KS3	GCSE	Total
2000	605,000	620,000	580,000	554,000	2.35M

A recent estimate by the BBC's Education correspondent, Mike Baker, of all examination scripts and items of coursework involved in the formal examination system – i.e. GCSE, AS-level and A-level - totalled 24 million per annum. In partial mitigation, or at least explanation, of the Edexcel debacle, he noted that the scale of the A-level enterprise had pretty much doubled over the course of one year, as a total of 815,000 A-level entries in 2001, were added to by another 850,000 AS-level entries (BBC News Education website 26/01/02).

Calculating the cost of this industry is by no means easy (despite the government's love affair with measurement and statistical 'transparency') since national testing is not identified separately in government expenditure figures. At the high end of an estimate, government statistics indicate that 7% of the annual primary school budget (£678M) and 8% of the annual secondary budget (£660M) is spent on "administration and inspection costs" including "central government expenditure on qualifications" (DfES 2001c, p. 17). Of this £105M was spent on OfSTED inspections in 2000 (p.12). At the low end of an estimate we can also note that the Qualifications and Curriculum Authority, now responsible for overseeing the National Curriculum and National Assessment arrangements, had a budget of £68M in 2000, of which £25M was identified as serving "Key Objective 4: Secure a rigorous, consistent and fair system of assessment" (QCA 2001). This low figure will certainly include QCA staff salaries and may include some of the payments to consortia for producing and marking the tests. So a reasonable guesstimate of the direct cost of national testing might be somewhere between £25M and the £100+M it takes to fund OfSTED. Of course, none of these figures take into account the indirect costs on local authority and teacher time; nor do they include the cost of GCSE examining, which would have to be calculated from the individual fee income received from candidates by each Examination Board.

Summary

So where does this leave us? Before going on to outline an alternative approach to using assessment to raise standards it is important to try to summarise the costs and benefits of a national curriculum and assessment programme. The positives can be summarised in terms of clarity, direction and organisation. Britain, or to be more accurate its constituent parts, and particularly England, now has an integrated education system in a way which it simply did not prior to 1988. There are clear goals and the organisational means to achieve them. Research indicates that, the initial overload having been sorted out, teachers are now largely happy with the content of the National Curriculum, subject-by-subject, especially in the primary sector which was seen as in need of significant reform. They have also been reported becoming more confident over time and identifying benefits in more collaborative whole-school planning, rather than simply being left to 'get on with things' as they had always done in their own individual classrooms (Pollard et.al. 1994, Croll 1996 and Earl et al 2001). Attention to use of evidence and criteria, rather than 'intuition', when it comes to making classroom assessments and appraising pupil progress has also been noted as an outcome of national assessment (Gipps et. al. 1995).

But this refocussing of the system has come at significant cost and, in particular, policy has not been able to avoid the enduring threat of any narrowly-based testing programme – that of coaching for the test and concomitantly narrowing the curriculum to the small number of objectives that are amenable to large-scale paper-and-pencil testing. All international research evidence suggests that this is what does happen, would happen, and it has happened. Furthermore, England now uses such tests at more ages and stages than any other country and the pressures are beginning to show very obviously. Additionally, the issue of curriculum stagnation has not been addressed, and even the government's evaluation of the National Literacy and Numeracy Strategies notes:

will the government retain the energy and resources needed several years from now to continually update materials [and] improve on prescribed practices...? This seems an unlikely long-term direction and one that might paradoxically result in a culture of dependence at local level...

(Evaluation of NL&NS: Earl et.al.2001, p. xii)

What is to be Done?

Fortunately there are alternatives, and the time is now right to turn to them. Testing, or more broadly conceived, assessment, will always have a profound impact on the curriculum, teaching and learning; the point is to maximise the beneficial impact and minimise the unintended side-effects. This isn't easy and will always involve trade-offs and compromises, especially if the political imperative is to 'roll out' developments across a large-scale system. The scale and scope of any new approach to assessment will always threaten its quality. Nevertheless we do know things about the ways in which assessment can be used to underpin, rather than undermine, high quality teaching and learning which are worth stating.

First and foremost we must get away from over-reliance on a single high-stakes indicator of performance – i.e. test results and only test results. Single indicators are always open to manipulation and corruption, by actors practising to effect the indicator, rather than improve the quality of the underlying goal which it represents. All studies of the validity and reliability of assessment indicate that multiple measures derived from different sorts of assessments – essays, orals, practicals, projects and so forth – produce more reliable results (cf. Cresswell 2000). Essentially this is an issue of sampling. Educational objectives are far more complex than can be reliably tested by one instrument on one occasion – different instruments or approaches must be used on several occasions. Additionally, by including multiple approaches to assessment in policy, our focus on the variety of educational objectives which we are pursuing will be reinforced rather than undermined.

Clearly however, large-scale multiple-goal, multiple-approach testing is not logistically feasible across three age cohorts at 7, 11 and 14 (four if we include GCSE). The original Key Stage 1 SATs, Standard Assessment Tasks as they were initially called, tried to operationalise such an approach but failed because it was over-ambitious and misunderstood by teachers still struggling to come to terms with the whole thrust of National Assessment. With hindsight it might be argued that teachers shot themselves in the foot by resisting more complicated forms of assessment, and in particular by arguing for external marking on the grounds of excessive workload, but nevertheless it is clear from experience here and in the United States that implementing large-scale centrally-controlled 'authentic assessment' as it known there, is extremely difficult (Torrance 1993, 1995; Khattri et.al. 1998).

Instead, in practical terms, implementing multiple approaches to assessment means resurrecting school-based assessment across the curriculum, or 'teacher assessment' as the original National Assessment policy report called it (TGAT 1988). In other words we need to extend GCSE-type coursework assessment and practical assessment across the system.

Public confidence in such widespread use of teacher assessment would have to be ensured by supporting such assessment with national item banks of centrally developed tasks, to be used towards the end of each Key Stage of the National Curriculum, as needed by classroom teachers. Use of the items would also have to be supported, in the first instance at least, by an in-service programme as ambitious as the National Literacy and Numeracy Strategies. Teachers would both use the tasks and, in time, develop them and devise others as their skills increased. Improvement of the items and of classroom assessment would be an interactive mutually reinforcing process. Elements of such a system of 'assessment instruments on demand', available from a centrally held item bank, already exist in Scotland. The main difference is that the Scottish item bank comprises simple test items for checking and confirming teachers' classroom judgements. My proposal is for more challenging tasks to be developed to broaden the scope and quality of teacher assessment.

In service support should include not only how to administer and mark tasks, but also how to analyse results, so that, in time, serious 'error analysis' as it is known in the literature, could be developed at school and consortium level, with LEA and HE support, to indicate what sorts of common misconceptions were being presented by pupils. Such work is already done for research purposes and should be turned into a resource for classroom teachers (cf. Williams and Ryan 2000, for a recent example of such work).

In order to address issues of scale and logistics, this renewed commitment to teacher assessment would also have to involve reducing the total amount of national testing as currently understood; cutting it back to, at most, Key Stage 2, age 11, only. Results could still be reported, to parents and via the school's annual report, but not via league tables, and effectively they would be the results of teacher assessment at ages 7 and 14, and possibly of teacher assessment combined with national tests at age 11. Aggregate reporting of results at LEA level might still be politically necessary at the end of KS2. In time GCSE could also be replaced by a similar form of centrally supported teacher assessment, with external accreditation being postponed to AS-level, as KS4 begins to be conceptualised as a genuinely 3-5 year programme (14-19), rather than assuming the traditional 'clean break' with a school-leaving examination at 16+.

The pattern of development would thus comprise:

- Central development of an item-bank of 'authentic' assessment tasks encapsulating meaningful and challenging curricular objectives, to test performance more validly and underpin high quality teaching;
- Used by teachers to structure their own classroom work towards the end of each Key Stage and inform their teaching more generally;
- In-service support to administer and mark tasks in similar fashion, leading to further development of local and possibly national items;
- In-service support to analyse patterns of pupil responses and particularly common errors and misconceptions.

The authentic assessment tasks which are devised for the developing national item bank could also be used periodically under more strictly controlled conditions, by the task developers, on a sample basis, to monitor performance in similar fashion to the American NAEP and the old APU. This would add to public confidence in results over time and alleviate political anxieties.

Formative Classroom Assessment

Just as important however, if not more so, is to continue to develop teachers' formative assessment skills at classroom level. This is where good assessment practice can make a positive difference to the curriculum experiences and quality of learning of individual pupils. Poor classroom assessment can confuse and intimidate pupils. But clear communication of the educational purpose underlying classroom tasks, the criteria by which good quality work will be judged, and, in turn, feedback to pupils about their work focussed on those criteria, will bring real gains in terms of educational standards. We have known this for some time. It is one of the clearest findings deriving from a great deal of classroom research and studies of the impact of assessment on learning (Crooks 1988, Black and William 1998, Assessment Reform Group 1999, Torrance and Pryor 1998, 2001). Yet it is being ignored by policymakers and, often, too narrowly interpreted by teachers in terms of very short-term goals of target-setting and raising test scores. This is because the development of high quality classroom assessment practices takes time, and is less immediately visible than a Literacy or Numeracy Strategy which can be manifest in materials and set procedures. Good formative classroom assessment is about careful observation of and *response* to pupils' efforts. It is much more difficult to encapsulate in a deliverable formula. Yet its core constituents are clear enough, and, taken together with the production of high quality assessment tasks for teachers to use as required, it could and should form the basis for a very different approach to the use of assessment to raise standards. Good formative assessment requires that:

- teachers are clear about their curriculum goals, shorter term learning intentions and the purpose of classroom tasks in relation to those learning intentions;

- communicate these intentions and the purpose of tasks to pupils - i.e. what they want pupils to do and why they want them to do it - *task criteria*;
- similarly, communicate to pupils what it means to do tasks well - *quality criteria*;
- make comments, mark work and give feedback relating to these criteria - indicating positive achievement as well as what and how to improve;
- be alert to unanticipated learning outcomes and encourage them when encountered - i.e. be alert to the possibilities for *divergent* as well as *convergent* assessment.

This is a tall order, especially in current circumstances, but it absolutely must become the focus of future development, if schools, the Exam Boards, the whole target-setting system, but more importantly, our children, are not to collapse under the cumulative weight of a too narrowly conceived testing regime. I started this paper with an engineering metaphor - testing the system to destruction. A crucial way of avoiding this is not just to reduce the load, but also to spread it and approach the problem from a different angle and with a different solution. Wales and Northern Ireland have dropped the publication of league tables; Scotland only ever reported results to parents and already employs a version of the item banking described above; further afield the Americans and others are exploring ways in which more flexible, authentic, approaches to assessment can be operationalised across a school system. It really is about time England did the same. Can testing really raise educational standards? No, but good quality assessment can, and the sooner the government understands this, the better.

Correspondence:

Professor Harry Torrance
 Head of Research
 Institute of Education
 Manchester Metropolitan University
 Didsbury Campus
 799 Wilmslow Road
 Manchester
 M20 2RR
 UK.

Email: h.torrance@mmu.ac.uk

References

- Atkin M. (1979) 'Educational Accountability in the United States' Educational Analysis, 1, 1, 1979, 5-21
- Assessment Reform Group (1999) Assessment for Learning, University of Cambridge School of Education
- Black P. and William D. (1998) 'Assessment and Classroom Learning' Assessment in Education, 5, 1, 7-74
- Cresswell M. (2000) Research Studies in Public Examining, Guildford, Associated Examining Board
- Croll P. (Ed. 1996) Teachers, Pupils and Primary Schooling: continuity and change, London, Cassell
- Crooks T. (1988) 'The impact of classroom evaluation on students' Review of Educational Research, 5, 4, 438-481

- Dearing R. (1993) The National Curriculum and its Assessment: Final Report, London, School Curriculum and Assessment Authority
- Department of Education and Science (1977) Education in Schools: A Consultative Document London, HMSO (Cmnd.6869)
- DES (1980) Statistics of Education: School leavers CSE and GCE results: England 1980, London, DES
- Department of Education and Science (1987) The National Curriculum 5-16 a consultation document, London, DES
- Department for Education (1992) Testing 7 Year Olds in 1992: results of the national Curriculum assessments in England, London, DfE
- Department for Education and Employment (2001) Statistics of Education: National Curriculum Assessments of 7, 11 and 14 year olds in England - 2000, London, DfEE
- DfES (2001a) Statistics of Education: Public examinations GCSE/GNVQ and GCE/AGNVQ in England 2000, London, DfES
- DfES (2001b) GCSE/GNVQ and GCE A/AS/VCE/Advanced GNVQ results for young people in England 2000/2001 (Provisional), London, DfES
- DfES (2001c) Education and Training Expenditure Since 1991-92, London, DfES
- DfES (2002) 14-19: Extending opportunities, raising standards
<http://www.dfes.gov.uk/14-19greenpaper/>
- Earl L. et. al. (2001) Watching and Learning 2: OISE/UT Evaluation of the Implementation of the National literacy and Numeracy Strategies Toronto, Ontario Institute for the Study of Education (for the Standards & Effectiveness Unit of the DfES; ref: DfES 0617/2001)
- Gipps C. et. al. (1995) Intuition or Evidence?, Buckingham, Open University Press
- Joseph K. (1996) 'The Role and Responsibility of the Secretary of State' in DES Better Schools: evaluation and appraisal conference, London, HMSO
- Khattri N. et.al. (1998) Principles and Practices of Performance Assessment, NJ, Lawrence Erlbaum Associates
- Kingdon M. and Stobart G. (1988) GCSE Examined, London, Falmer Press.
- Klein S. et.al. (2000) 'What do test scores in Texas tell us?' Education Policy Analysis Archives 8, 49, 2000; <http://epaa.asu.edu/epaa/v8n49>
- Linn R. (2000) 'Assessments and Accountability' Educational Researcher, 29, 4-16
- Linn R. (2001) The Design and Evaluation of Educational Assessment and Accountability Systems: CSE Technical Report 539, CRESST, UCLA.
- McNess E. et. al. (2001) 'The Changing Nature of Assessment in English Primary Classrooms' Education 3-13, 29, 3, 9-16
- OfSTED (2001) The Annual Report of Her Majesty's Chief Inspector of Schools, 2000/2001
<http://www.official-documents.co.uk/document/deps/ofsted/hc500/352-02.htm>
- Osborn M. et. al. (2000) What Teachers Do: changing policy and practice in primary education, London, Continuum

- Pollard A. et. al. (1994) Changing English Primary Schools, London, Cassell
- Popham J. (1987) 'The merits of measurement-driven instruction' Phi Delta Kappan, 68, 679-682
- Qualifications and Curriculum Authority (2001) Annual Report,
www.qca.org.uk/annual_report/resourcing.asp
- Resnick L. & Resnick D. (1992) 'Assessing the thinking curriculum' in Gifford B. & O'Connor M. (Eds 1992) Future Assessments: Changing Views of Aptitude, Achievement and Instruction, Boston MA, Kluwer
- Shepard L. & Smith M. (Eds 1989) Flunking Grades: research and policy on retention, London, Falmer
- Task Group on Assessment and Testing (1987) National Curriculum: Task Group on Assessment and Testing: A Report, London, DES
- Torrance H. (1981) 'The origins and development of mental testing in England and the United States' British Journal of Sociology of Education, 2, 1, 45-59
- Torrance H. (1993) 'Combining measurement driven instruction with authentic assessment: the case of national assessment in England and Wales' Educational Evaluation and Policy Analysis, 15, 1, 81-90
- Torrance H. (1995) Evaluating Authentic Assessment: problems and possibilities in new approaches to assessment, Buckingham, Open University Press
- Torrance H. (1997) 'Assessment, accountability and standards: using assessment to control the reform of schooling' in Halsey A.H. et. al. (Eds 1997) Education: Culture, Economy and Society, Oxford, Oxford University Press
- Torrance H. and Pryor J. (1998) Investigating Formative Assessment: teaching learning and assessment in the classroom, Buckingham, Open University Press
- Torrance H. and Pryor J. (2001) 'Developing Formative Assessment in the Classroom: using action research to explore and modify theory' British Educational Research Journal, 27, 5, 615-631
- Williams J. and Ryan J. (2000) 'National Testing and the Improvement of Classroom Teaching: can they co-exist?' British Educational Research Journal, 26, 1, 49-73