Standards in English Primary Schools:

are they rising?

A contribution to the debate from the Association of Teachers and Lecturers
written and researched by Colin Richards
ATL. Publications and a whole lot more.

As the leading union that promotes education, ATL exists to protect and improve the professional interests of teachers, lecturers and all staff involved in providing education. We support our members with unrivalled research, publications and advice, in all aspects of their professional and working lives.

ATL. There to help.

ATL provides expert legal advice and representation for all members. A network of experienced solicitors and caseworkers are ready to help members with any legal problem they may encounter at work.

ATL. An authoritative voice.

As a highly influential trade union and professional association, ATL can be relied on to promote the interests of its members and the cause of education at every opportunity.

Decision-makers listen to ATL's views because of its constructive approach, based upon the experiences of members – real practitioners, reporting from the front line of the profession.

ATL. Not a member?

To be part of the Association that puts education first and for complete protection and peace of mind in the classroom, join the Association of Teachers and Lecturers today. If you're a newly qualified teacher, membership is FREE for the first year. All other new members benefit from a 50% discount off membership for the same period. You can join online by setting up a Direct Debit* at www.atl.org.uk. Alternatively phone us on 0845 057 7000**

For a free copy of the ATL publications catalogue, please contact the ATL despatch line on 0845 4500 009.

* terms and conditions at www.atl.org.uk
** local rates apply
CONTENTS

1. Preface 4

2. Foreword 5

3. Introduction 6

4. Distinctions 8
   4.1 Standards 8
   4.2. Rising/falling standards 9
   4.3 Improving/deteriorating standards 10
   4.4. Performance 10
   4.5. Rising/improving performance 11

5. Standards: the evidence 12

6. Performance: the evidence 14
   6.1 Introduction 14
   6.2. Review of evidence related to performance 14
      6.2.1 Data from statutory tests 14
      6.2.2 Data from statutory teacher assessments 19
      6.2.3 Inspection data from Ofsted 20
      6.2.4 QCA test data related to the national strategies 22
      6.2.5 Test data from major research projects 23
      6.2.6 Other data 25

7. Overview and policy implications 27
   7.1 Standards and performance: an overview 27
   7.2. Policy implementations 29
      7.2.1 Keeping standards under review 29
      7.2.2 Monitoring performance against standards over time 30

8. Conclusion 31

9. References 33
Preface

This paper argues that discussion about changing standards in English primary schools has been characterised by inadequate conceptualisation that has ignored a number of crucial distinctions. These distinctions are explored in relation to national test data and data from research projects. The importance of expert judgement is stressed. Due to conceptual and other inadequacies no firm conclusions are possible in relation to the question posed in the title of the paper. However, some tentative intimations are suggested. The paper concludes by making policy recommendations for keeping standards under review and for monitoring performance against standards over time.

Professor Colin Richards

Colin Richards is Professor of Education at St Martin’s College and visiting professor at the Universities of Leicester, Warwick and Newcastle and at the National Institute of Education in Singapore. He was a primary class teacher, deputy headteacher and lecturer in education before becoming an HMI in 1983. After inspecting schools in the North East of England he became Staff Inspector (Curriculum) in 1987. From 1994 to 1996 he was Ofsted’s Specialist Adviser in Primary Education. While an HMI he wrote or edited a number of official publications related to the curriculum and to primary education. Since leaving HM Inspectorate he has maintained a keen interest in the issues of standards, the primary curriculum and school inspection. He has published widely and is a frequent contributor to the national press, particularly the *Times Educational Supplement* and *Education Journal*. As well as being a paid-up member of ATL and a small school enthusiast he is chair of governors of a Cumbrian secondary school. He greatly treasures the epithet ‘an old-fashioned HMI’ bestowed on him maliciously by a former chief inspector of schools.

Acknowledgements

The author would like to acknowledge the help of the following colleagues who made valuable comments on earlier drafts of this paper: Norman Thomas, Peter Tymms, Neil Simco, Christopher Schenk and, as ever and always, Sheila.
Foreword

This publication could not have been written at a better time. As teachers know to their cost, much of the current ‘debate’ about school standards sheds more heat than light. Standards have become a political football. The ritual controversy about whether standards are rising or falling that accompanies the annual publication of school league tables certainly keeps the media busy. It does not, however, even begin to get to the heart of what we mean by standards. Standards in English Primary Schools: are they rising? is a welcome and long overdue antidote to much of the sloppy and muddled thinking in which debates about school standards are too often couched.

Let us take an example. Ofsted – the Office for Standards in Education – tells us that, overall, standards in teaching are improving, in some cases significantly so. Almost in the same breath, we hear from the Qualifications and Curriculum Authority (QCA) (whose ‘strapline’ is ‘Guarding Standards’) that standards in National Curriculum tests have plateaued. How can it be that teaching is improving yet test results are not? And only a year or so ago I challenged the head of the government’s former Standards and Effectiveness Unit when he stated unequivocally (and without a shred of evidence) that standards had not improved in schools in England for several decades until New Labour came to power. Could it perhaps be the case that we are not talking about the same standards – and if so, why is that not made clear to schools, teachers and, perhaps most important of all, to the public?

This publication is an important reminder of the complexities that are embedded in the seemingly straightforward question: Are primary school standards rising? In addressing this question, and in offering some thoughts as to what the answer might be, Colin Richards takes a scholarly, deeply analytical, evidence-based approach. In a tightly-argued analysis of the key issues, he reminds us of the importance of choosing our words with care. Standards, for example, are criteria used to assess the quality of pupils’ performance. It is therefore important to distinguish between ‘standards’ and ‘performance’ in relation to those standards, and to take a critical, informed look at the publicly available evidence on both standards and performance. For readers with little time to spare, each section ends with a helpful set of ‘take-away’ messages, drawing together the main conclusions.

Colin Richards brings to the debate an enthusiasm for the subject, a wealth of information drawn from a lifetime in education and, importantly, someone who has always shown the highest regard for the teaching profession, and for the children and young people the profession serves.

This publication will be pertinent to all with an interest in the ‘standards debate’: teachers, headteachers and the wider school workforce, school governors, students of education and, of course, policy-makers and politicians – without whom the current debate about school standards would barely have begun.

Dr Mary Bousted, ATL General Secretary.
Introduction

There is a great deal of confused thinking about the issues posed in the title of this paper. That confused thinking, along with the paucity of reliable data, results in periodic ‘moral panics’ which hinder, rather than aid, the provision of effective advice to policy-makers and the conduct of a sensible professional debate.

This paper focuses on primary education in England (especially Key Stage 2) and on the period since 1995, the first year in which data based on tests conducted in all English maintained primary schools were made publicly available. That period has seen very considerable, some would say ‘remarkable’, rises in statutory test scores at the end of Key Stage 2, particularly in the period 1995-2000. These rises have had both political and educational significance. They have been, and still are, cited as one of New Labour’s major successes – providing evidence of the government’s commitment and energy in improving public services through the raising or ‘ratcheting-up’ of standards. In educational circles they have been used to justify the government’s interventionist and prescriptive policies for primary education and have been cited as justification for similar initiatives in Key Stage 3. There is no doubt that national testing and the importance attached to the results have had a major influence on policy and practice in schools. As ‘high stakes’ tests, they provide many types of influence. They influence what is taught, how it is taught, how it is assessed and how teachers and children are motivated (or de-motivated). They have been, and still are, used to reinforce particular aspects of the curriculum and to signal that others are less important or even marginal.

Assessment experts are well aware of the technical difficulties in attempting to answer the question posed in the title of this paper. There is general acceptance that ‘assessing standards over time is never easy’ and that ‘despite the assessment reforms, it remains difficult to measure changes in national standards with complete confidence’ (Tymms and Fitz-Gibbon, 2001 p.158).

Not all, however, go as far as Wiliam (2001) in arguing that ‘the whole idea of measuring standards over time in any real sense is nonsense – not just with National Curriculum tests, but with any tests of achievement (p.11). However, assessment experts are far less aware of conceptual distinctions that need to be made if any initial sense is to be made of the question.

So what are ‘standards’? Do ‘rising standards’ automatically imply ‘improving standards’? What is the relationship between test scores and ‘standards’? What does the evidence suggest? Where is the role of expert judgement in this contentious, confused area?
This paper attempts to clarify thinking by making a number of crucial distinctions, to consider evidence, to offer tentative explanations for findings and to explore the inevitable controversies, unreliabilities and uncertainties surrounding the vexed issue of standards.

In brief

- There is a great deal of confused thinking and talk about standards
- Assessment experts agree that it is very difficult to assess standards with confidence.
Distinctions

It is crucially important to distinguish between ‘standards’ on the one hand and ‘performance’ in relation to standards on the other (Note 1). This is not just a minor matter of phraseology; it has major implications for the way in which the question in the title of the paper can, and should, be answered.

4.1 Standards

Standards are ‘benchmarks… criteria whereby one assesses or evaluates the quality of a particular activity or process. And that quality must depend upon the identification of the purpose of the activity – upon the values which are embodied within it’ (Pring, 1992 p.20). Standards by which an educational activity or performance (in our case performance nationally in English primary schools since 1995) is judged needs to be distinguished from how far the performance itself meets those standards.

Where are standards to be found? They are implicit in any test or form of assessment but are usually left implicit. Each of the tests referred to later in this paper embodies such standards but the nature of such standards is not discussed within the assessment literature nor is the relationship between the standards of the various tests and the nationally-derived standards ever reported on in any detail.

England does have nationally derived standards (or benchmarks) – they are embodied in National Curriculum level descriptions. What is their basis? How have they been derived? The procedural answer is that they originated a decade or so ago in the deliberations of government-appointed subject working parties, were developed in their current form following the Dearing Review in 1993, and were revised for Curriculum 2000 by officials and working parties of the Qualifications and Curriculum Authority (QCA) following ‘consultation’ with teachers and other interested parties. The level descriptions have been derived pragmatically from a consideration of ‘good practice’ (value-laden by definition as if good practice wore its credentials on its face!) and from the beliefs, practices and assumptions of officials and working party members. The existence of expert judgement among the officials and members has been assumed, not demonstrated; the nature of that judgement and the means by which it was secured or validated have not been discussed in the public arena. Yet as will be argued later, ‘expert judgement’ is of crucial significance in the establishment of standards, as is acceptance of the legitimacy of that judgement by all stakeholders in the assessment process.

Two allied characteristics of the level descriptions need to be borne in mind. They are not tightly defined lists of performance criteria spelling out specific behavioural competences but are much more loosely defined descriptions embodying more general criteria. They capture knowledge of a general and sometimes abstract character and because of the non-technical ‘every-day’ nature of the language in which they are expressed they are inevitably imprecise
and capable of being interpreted differently (albeit within limits) by those who use them. Frater (1999) illustrates the same point graphically in relation to level descriptions for English:

‘the descriptions remain profoundly indeterminate. The crucial Level 4 for writing is a clear case in point:

“Pupils’ writing in a range of forms is lively and thoughtful. Ideas are often sustained and developed in interesting ways and organised appropriately for the purpose of the reader. Vocabulary choices are often adventurous and used for effect. Pupils are beginning to use grammatically complex sentences, extending meaning. Spelling, including that of polysyllabic words that conform to regular patterns, is generally accurate…” (QCA, DfEE 1999)

A range, lively, thoughtful, often sustained, interesting, appropriately adventurous, generally accurate, beginning to use – each is open to wide interpretation. Placing such value-laden terms at the heart of a level description blunts and softens exactly those hard edges that criterion-referencing was invented to provide.’

Very importantly level descriptions are in some sense arbitrary. They could have been conceived and expressed differently; different people could have come up with different formulations. Those with a different view of a particular subject could have produced level descriptions embodying different standards. The arbitrary, potentially contentious, basis of these level descriptions needs to be borne in mind. However, nowhere in official documentation or in the research literature is there any acknowledgement of the arbitrary nature of the level descriptions used by the government and its agencies such as the Department for Education and Skills (DfES), Ofsted or QCA.

To reiterate, in answering the question of whether standards are rising or improving the nature of those standards needs to be made clear.

4.2 Rising/falling standards

Standards can be said to be ‘rising’ or ‘falling’ if over time the criteria embodied in level descriptions are modified to provide greater, or lesser, intellectual challenge. Such modifications may be deliberately designed to have a particular effect; they may not be. But any modifications to the wording of such descriptions imply changes, however slight, to standards. Such wording has changed over time – compare for example the wording of the first set of level descriptions for English and for mathematics (following the Dearing Review of 1993) with those in Curriculum 2000. Whether such changes constitute a rise, fall or neither in standards of intellectual challenge can only be determined by expert judgement. It is an interesting educational and political question how that ‘expert judgement’ is secured (see section 7.2).
4.3 Improving/deteriorating standards

Standards can be said to be ‘improving’ or ‘deteriorating’ if over time the criteria embodied in level descriptions are judged to be more or less valuable by those exercising expert judgement. The latter might judge, for example, that one or more of the criteria have become less relevant over time or are inappropriate in the light of recent research evidence. In contrast, revision to such criteria might be judged as resulting in improving standards if the modifications are believed to be more suitable to changed circumstances. In the former case standards might be judged to have deteriorated somewhat, whilst in the latter case to have improved. But such judgements are value judgements and need to be made by expert opinion.

4.4 Performance

The literature and debate on standards usually fails to distinguish between standards as criteria and performance related to those standards. For example, in successive annual reports Ofsted has discussed children's performance in Key Stages 1 and 2 based on analyses of national test data without ever discussing the nature of the standards against which that performance is being assessed. Tymms (2004) poses the question ‘Are standards rising in English primary schools?’ but concentrates entirely on performance – neglecting the more fundamental issue of the criteria (or standards) to which that performance is related.

The means used to assess performance are invariably tests of one sort or another – national tests provided by the QCA, the Performance Indicators in Schools Project (PIPS) ‘bespoke’ tests (Tymms and Coe, 2003), Leverhulme’s numeracy tests ‘derived from instruments developed from earlier research’ (Brown et al. 2003); Davies and Bremer’s use of publicly available ‘standardised tests’ and Trends in International Mathematics and Science Study’s (TIMSS) tailor-made tests (Ruddock, 2004). Tellingly, in discussion about ‘rising’ or ‘falling’ standards no reference is made to statutory or non-statutory teacher assessments and how these have changed over time. In its reporting of ‘standards achieved’ the annual reports of Her Majesty’s Chief Inspector of Schools (HMCI) do not discuss teacher assessments nor do they report inspectors’ own judgements of performance against standards – a point discussed in section 6.2.2. In the current debate test performance is erroneously equated to standards.

Assuming, for the sake of argument, the validity of using tests as measures of performance against national standards (see Davies, 1999, for a different view) there are a number of different ways in which this could be attempted:

1. Children’s performance against the standards embodied in level descriptions could be assessed over time using the same tests administered in the same way. This has not been attempted in England in connection with the National Curriculum due to the ‘high stakes’ nature of national testing and the unavoidable temptation for teachers and schools ‘to teach to the test’.
2. Children’s performance against the standards embodied in tests not directly designed to relate to level descriptions could be assessed over time using the same tests administered in the same way. This is the approach used by Leverhulme, PIPS, Davies and Brember and TIMSS.

3. Children’s performance against the standards embodied in level descriptions could be assessed year-on-year by different tests, but with the tests in successive years being made as comparable as possible. This is the approach adopted by the QCA and its predecessor, the School Curriculum and Assessment Authority (SCAA).

It is important to remember that only two of the three approaches have provided test evidence. These are based on different conceptions of ‘standards’ and it is important to distinguish between their different conceptions of performance against standards.

4.5 Rising/improving performance

The distinctions made earlier between ‘rising’ and ‘improving’ standards (sections 4.2, 4.3) do not apply in the context of performance ‘measured’ on scores derived from the same tests administered at different times. Provided that the same tests are used and that they reflect criteria (or standards) regarded by users as valuable then any increase in test scores equates to a rise in performance and thence to an improvement. However, it is problematic how far changes in performance can be equated with rises and hence improvements where (as in national testing) different tests are used in successive years and where their relationship with standards is unclear (Hilton, 2001).

In brief

[•] It is crucially important to distinguish between ‘standards’ and ‘performance’ in relation to those standards.

[•] Standards are criteria used to assess the quality of children’s performance.

[•] In England the criteria embodying standards are embodied in National Curriculum level descriptions.

[•] Standards rise or fall if the criteria in the level descriptions or tests become more or less challenging.

[•] Standards improve or deteriorate if in the view of ‘experts’ they are more or less valuable as circumstances change.

[•] Children’s performance in relation to those standards is usually measured by tests.

[•] Performance rises or falls, improves or deteriorates, if children’s scores on these tests increase or decrease.
Standards: the evidence

What evidence is publicly available as to how far the standards embodied in level descriptions have been reflected in the tests used by either official agencies such as the QCA or by researchers?

Neither the QCA nor its predecessor the SCAA has published any detailed analyses of the relationship between test items and level descriptions. To illustrate, there is no published information about:

(a) how far any one test (say in reading at the end of Key Stage 2) reflects particular aspects of the level descriptions for reading

(b) whether successive tests reflect the same aspects over time

(c) whether successive tests retain the same balance between items related to aspects of level descriptions

(d) which aspects of the level descriptions are not reflected in test items.

Equally there is no published work by researchers examining such questions, with the exception of the single study by Hilton (2001) (cited previously), which provides evidence in relation to (c). Though not providing an item-by-item comparison of test items with level descriptions, Hilton uses the QCA’s own categories of reading skills in her analyses of test items in the Key Stage 2 tests for reading in 1998, 1999 and 2000. Hilton argues that ‘What is essential... is that the analytical categories... should remain constant in any analysis of subsequent tests. Yet each year from 1998 the analytical categories that the QCA use to discuss the questions and children’s responses have changed. It is, however, only by using a constant set of categories that a genuine year-by-year comparison can be achieved’ (p.5).

If Hilton’s analysis is accepted then the standards embodied in those tests changed over time, therefore rendering it impossible to compare children’s performance year-on-year in a meaningful way.

It is surprising that the academic community has not conducted similar analyses of other national tests. Presumably the test developers employed by the QCA have undertaken analyses in relation to (a) to (d) above, but if they have why have their findings not been publicly reported and debated so the teaching profession and others could be satisfied that the standards were not being changed (consciously or unconsciously)?

Some of the same criticisms can be levelled at those researchers who have used tests (not chosen for their direct relationship to the level descriptions) to assess children’s performance over time. There has been no published information as to the criteria embodied in the tests used and how far they relate to the criteria embodied in level descriptions. Instead there are only very general statements. For example Tymms and Coe report that ‘the PIPS project has
bespoke tests’, although two were taken off the shelf (Tymms and Coe, 2003 p.645) but do not disclose the criteria governing the choice of items in those ‘bespoke’ tests. Again Brown et al. (2003) indicate in their Leverhulme numeracy project that ‘a sequence of tests, one for each year group, was used, which were derived from instruments developed from earlier research. A few new items were trialed and added to further improve the match with the National Numeracy Strategy’ (p.660), though these tests did not relate to areas such as space and shape and data handling – important aspects of the level descriptions for mathematics. Similarly, in their summary of a major international comparative study Ruddock et al. (2004) simply state that ‘the intended curriculum for England in science and mathematics matched the topics tested in TIMSS well’ (p.11).

To reiterate, there is no readily available published evidence as to how far national standards, as embodied in level descriptions, have been reflected in the tests used by national agencies or researchers. Therefore it is impossible to say with any certainty whether any changes in performance as ‘measured’ by tests reflect changes in performance against national standards.

In brief

- **There is no published evidence on how far national standards (as embodied in the level descriptions) have been reflected in the national tests or in the tests used by researchers to assess children’s performance over time.**
Performance: the evidence

6.1 Introduction

In this section evidence related to children’s performance over time is reviewed from five main sources and two minor ones. Two preliminary points need bearing in mind. Firstly, though not discussed in detail here, there are numerous controversies among assessment specialists over the use and interpretation of test data, including whether:

(a) it is possible to measure (as opposed to appraise) performance in relation to the kinds of standards embodied in National Curriculum level descriptions

(b) it is possible through testing to assess ‘connected’ knowledge and understanding (as opposed to the ‘thin’ knowledge required to answer test questions) (Davies, 1999)

(c) the most important aspects of particular subjects are, or can be, subject to meaningful testing

(d) the particular tests used can, and do, reliably and validly assess performance in relation to particular level descriptions

(e) the particular tests can be marked, and the results reported, fairly and consistently, especially where criteria are not precisely defined and the marking is inevitably subjective (Richards, 2001).

The second and very important preliminary point is that the data resulting from statutory tests relate to only two subjects at Key Stage 1 and to three subjects at Key Stage 2 – and only to two year groups. Despite, for example, the eloquent well-argued plea in All Our Futures (DfEE, 1999) for the role of the arts and other aspects of creative education, these have not featured in national assessment. It needs to be remembered that children’s performance in eight out of the 11 subjects of the mandatory curriculum has not been addressed (fortunately) through testing.

6.2 Review of evidence related to performance

6.2.1 Data from statutory tests

The most widely quoted data related to children’s performance in primary schools are the official statistics compiled by the DfES relating to the percentage of children achieving each of the prescribed levels in the national tests for mathematics, English and science. The key statistics are the percentage of children reaching a so-called Level 4 at the end of Key Stage 2 and, of somewhat lesser importance, the percentage reaching a so-called Level 2 in English and mathematics at the end of Key Stage 1. Tymms (2004) neatly summarises the arrangements for Key Stage 2 testing:
The test material is the responsibility of the QCA… Two sets of new tests are produced each year with one being kept in reserve in case of unexpected problems. Cut-scores are set to identify which pupil has attained which level after the national data become available. The tests are the subject of stringent security, and checks are made to ensure that proper procedures are followed in schools. The tests are marked externally and the scripts returned to schools after marking’ (p.479).

Figures 1 and 2 provide a summary of the data since 1996. The results for reading and writing at the end of Key Stage 1 show a small but steady rise in performance from 1996 to 2001 followed by a plateauing of results in reading and small variations in performance (including falls in 2003 and 2004) for writing. Scores in mathematics follow a similar pattern to those for reading – a somewhat surprising finding since the pattern of results over time in two very different subjects might be expected to vary.

**Figure 1: Percentage of seven-year-old pupils achieving Level 2 or above in reading, writing and mathematics tests/tasks at Key Stage 1**

The results for English at the end of Key Stage 2 show a substantial steady rise from 1996 to 2001, and followed by a plateauing of results in 2002 and 2003 by a slight rise in 2004. The results for mathematics show a not too dissimilar pattern – a steady rise from 1996 to 2000 except for a ‘blip’ in 1998 – probably occasioned by the introduction that year of an oral test (with a consequent change in the standards being assessed). The mathematics results tend
to plateau from 2000 to 2003 and then rise very slightly in 2004. The results for science show a similar pattern to those for English, with a steady rise to 2001 followed by a plateauing thereafter. Again the similar pattern of the results in the three very different subjects is surprising, particularly when one (science) had not been subject to national strategy while the others had.

Figure 2: Percentage of 11-year-old pupils achieving Level 4 and above in English, mathematics and science tests at Key Stage 2

The Key Stage 2 data have been widely cited by official agencies as major indicators of the success of government policies, especially those related to literacy and numeracy in primary schools. This interpretation has been backed up by two evaluation studies. In their analysis of national tests over time commissioned by the QCA, and despite having some reservations about test standards (discussed in the following paragraph), Massey et al. (2003) suggest that the National Curriculum and its associated testing regime have been successful in ‘levering up’ standards. Likewise in their evaluation of the National Literacy and Numeracy Strategies Earl et al. (2003) claim that the test data show a significant rise in attainment from 1997 to 2002. While acknowledging the difficulty of drawing conclusions about the effects of the strategies on pupil learning they report uncritically and without further reservations that ‘attainment on the government’s key measures rose significantly’ (p.3).
Children's performance on national tests *could* well have improved significantly since 1996 but before accepting this view other explanations need to be explored. It needs to be stressed at the outset that there is no evidence for the cynical view that the QCA has quietly, unobtrusively and consciously modified their standard-setting procedures to ensure a year-by-year rise towards the targets set by the government. There is, however, evidence (Massey et al. 2003) that, unlike in the Key Stage 1 tests and in the Key Stage 2 tests for mathematics ‘there may have been a shift towards more lenient standards in Key Stage 2 English test thresholds between 1996 and 1999’ – suggesting that the ‘national results might over-estimate the rate of progress’ (p.216) and so produce an ‘illusory’ rise.

Part of the explanation for the rise in children’s performance could be accounted for by problems with the mechanisms used by the QCA. For a number of years after 1995 the Authority’s approach to ensuring comparability from one test to the next, in terms of the National Curriculum levels awarded, involved four procedures:

1. the use of the same anchor test each year against which the results of any new test were checked
2. giving a sample of pupils taking the ‘real’ test in a particular year with the following year’s test so that equivalent scores could be estimated
3. the use of expert judgment in scrutinising scripts
4. the use of expert judgment in examining the questions set (Quinlan and Scharaschkin, 1999).

Tymms and Fitzgibbon (2001) suggest that in the context of high stakes testing resulting from the setting of national targets for English and mathematics, five mechanisms might be responsible, at least in part, for the reported rise in children’s performance up to 2000-01:

1. They argue that ‘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 is the anchor test’ but ‘this is not used to check standards across the years. It is used only to check standards from one year to the next’ (p.162);
2. In procedure (b), referred to in the previous paragraph, tests on next year’s test are equated to results on this year’s ‘real’ test even though the former are taken in very different ‘adrenaline-free’ circumstances;
3. Decision-making by examiners about cut-off marks for particular levels (especially level 4) might be influenced by knowledge of the impact they will have on primary schools’ progress towards the published high stakes targets;
(4) Markers might look more kindly on borderline cases as the stakes are raised year-on-year;

(5) The return of scripts to schools might increasingly result in schools challenging the marking and a consequent reaction among markers to give pupils the benefit of the doubt.

In a later paper Tymms (2004) argues that sometime after 2000 the QCA tightened its procedures and in relation to (1) took care to set cut-off scores on the basis of data collected not just in the previous year but over several years. He implies that this change might account at least in part for the levelling off of test data in recent years.

Such technical issues could be part of the explanation for the spectacular rise in test scores in the latter half of the 1990s, and in as far as they are valid could cast doubt on the extent of the supposed rise in performance.

However, other factors could, and almost certainly did play a part in helping produce that rise. The most important was the pressure on primary schools to improve their Key Stage 2 test results in English and mathematics, especially following the introduction of the National Literacy and Numeracy Strategies and the setting of national targets. Understandably schools focused more closely on those subjects, especially after 1998, when the Secretary of State lifted the statutory requirement for schools to follow the Key Stage 1 and 2 programmes of study in the non-core subjects. This was reinforced by Ofsted inspections which focused almost entirely on the tested subjects to the detriment of the rest of the curriculum. There was an undoubted rise in the extent of test preparation undertaken in primary schools, especially in Year 6. Children were given more instruction in test techniques as well as more practice per se in taking tests. Supporting this line of argument are the comments of Brown et al. (2003) about the mathematics and science results from 1997 to 2002: ‘The way the percentages grow and then plateau for both subjects, with the mathematical results improving less dramatically for science, even though there was no national science strategy, suggests that increasingly careful test preparation was the salient factor in improvement’ (p.669).

The fact that primary schools felt under pressure to improve their test results should not, however, detract from the fact that as ‘measured’ by national tests results did rise until 2000-01 in all three tested subjects even though part of that rise might be due to ‘technical’ factors associated with testing. What is more disputable is the educational significance of the rises. For example, it could be that rises in children’s performance, due in part to test preparation, may not represent permanent gains in learning and that performance might drop within months of taking the national tests. No research has been conducted, for example, on whether children at the end of their first term in Year 7 score as highly or higher on national tests than they did at the end of Year 6. There is a little research conducted at the end of Year 7 which lends credence to the notion that in a significant number of cases gains in learning represented by Year 6 scores may be short-lived (as is the case with most ‘high stakes’ tests);
Brown et al. (2003) report that ‘when pupils from the National Key Stage 3 Strategy pilot schools were tested at the end of year 7 in 2001, the percentage of the population reaching Level 4 appeared to be around ten percentage points lower than for the same pupils at the end of year 6...’ (p.669).

Again, to repeat an argument used extensively in this publication, rises in children’s performance as ‘measured’ by test scores do not necessarily involve a rise in performance against national standards unless it can be demonstrated that these standards are embodied in the same way and to the same degree in the successive tests. Without that assurance the educational significance of any rises is problematic and contentious.

6.2.2 Data from statutory teacher assessments

A second source of data relates to statutory teacher assessments since 1996.

The data (Table 1), here published for the first time, make interesting reading. At Key Stage 1 teacher assessments mirror the national test results in showing a small but steady rise in performance in English and mathematics from 1996 to 2001 with a levelling off thereafter. Both test data and teacher assessment data report a six per cent rise in the proportion of pupils achieving Level 2 in English at the end of Key Stage 1 and a seven per cent rise in mathematics. Teacher assessment data in science show a similar pattern, though with a five per cent rise. This close correspondence between test and teacher assessment data is not surprising given the procedures governing the assessment and reporting of children’s performance at the end of Key Stage 1.

Table 1: Teacher assessment results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>81</td>
<td>79</td>
<td>80</td>
<td>81</td>
<td>82</td>
<td>84</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>*</td>
</tr>
<tr>
<td>Mathematics</td>
<td>78</td>
<td>82</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>88</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>Science</td>
<td>84</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>90</td>
</tr>
</tbody>
</table>

* The combined English result was not available from the trial schools hence the 2004 results are not comparable to previous data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>57</td>
<td>60</td>
<td>63</td>
<td>65</td>
<td>68</td>
<td>70</td>
<td>72</td>
<td>73</td>
<td>72</td>
<td>74</td>
</tr>
<tr>
<td>Mathematics</td>
<td>54</td>
<td>60</td>
<td>64</td>
<td>65</td>
<td>69</td>
<td>72</td>
<td>74</td>
<td>74</td>
<td>74</td>
<td>75</td>
</tr>
<tr>
<td>Science</td>
<td>64</td>
<td>65</td>
<td>69</td>
<td>71</td>
<td>75</td>
<td>79</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>83</td>
</tr>
</tbody>
</table>
At Key Stage 2 teacher assessments rise steadily from 1996 to 2001 (plateauing thereafter) but at a rate of improvement considerably less than that recorded in the test results – 12 per cent more pupils achieving at least Level 4 according to teacher assessments compared with 18 per cent more on the tests. In mathematics there is a similar pattern with teacher assessments again recording a lower rate of improvement (14 per cent) compared with 17 per cent on the tests. In science a similar pattern occurs but at a higher level – with a 17 per cent improvement recorded by teacher assessments compared with a 25 per cent rise reported on the tests. It would appear that teachers of children at the end of Key Stage 2 have been rather more conservative in claiming large rises in performance than the government or Ofsted.

Tellingly trends in these statutory teacher assessments have never been published by the DfES or cited by Ofsted when discussing changes in performance – despite the ‘official’ line since the Dearing Review that such assessments should have the same standing as national test data.

6.2.3 Inspection data from Ofsted

Ofsted inspections ought to be a major source of data on standards and performance since by law its inspectors are required to report on four broad aspects of school education, the first of which is ‘the educational standards achieved in the school’. Since the beginning of the first round of primary inspections in 1994 Ofsted has collected a great deal of data on schools, including its inspectors’ judgements on standards. What does this data indicate about standards and performance against standards over time?

Throughout its existence Ofsted has failed to distinguish between ‘standards’ and ‘performance against standards’ (Note 2). Scrutiny of successive annual reports reveals a surprising (and problematic) variety of interpretations as to what Ofsted has taken ‘standards’ to be. In an early annual report (Ofsted, 1996) there are references to ‘standards’ based on judgements made in relation to ‘national expectations’ and to ‘pupils’ capabilities’. A later report (Ofsted, 1998) refers to ‘two separate judgements of standards achieved by pupils: attainment – how well pupils are achieving in relation to national standards or expectations; and progress – the gains pupils make in knowledge, skills and understanding’ (p.89). Still later, reports published in 2001, 2002 and 2003, provide yet another set of distinctions – between standards as ‘the educational attainment of pupils in relation to clear benchmarks such as National Curriculum levels’ and ‘how well the pupils achieve, taking account of the progress they have made and other relevant factors’ (Ofsted, 2001 p.89). The Ofsted subject reports for 2003-04 published as part of the latest Annual Report (Ofsted, 2005a) simply reports test results under the heading ‘Standards’ for the tested subjects (Note 3). The heading does not occur at all in the reports for the foundation subjects!

In consequence of these variations it is impossible to make a judgement of any changes in performance against standards over time based on the judgements reported publicly in HMCI’s Annual Reports. However, in individual school reports Ofsted has always required its
inspectors to make, and report, judgements about the school's performance in the range of subjects inspected. But Ofsted inspectors’ judgements of standards have never been aggregated and reported nationally to set alongside test data in mathematics, English and science or to provide an overview of performance in the non-tested subjects. Periodic publication of such aggregated judgements could have provided some evidence of changes in schools’ performance nationally over time. The ‘explanation’ offered by Ofsted for the absence of such judgemental data (Note 4) is unconvincing. It could be valuable to be able to compare over time, if only in a tentative, rough-and-ready way, inspectors’ judgements of how well schools were achieving in relation to national standards (embodied in level descriptions) in all the subjects of the primary curriculum. It would be particularly instructive to be able to set changes in these judgements alongside changes in national test data, which purportedly ‘measure’ pupils’ performance in relation to level descriptions. Why has Ofsted not provided the data necessary to make these comparisons? Is it because there have been marked discrepancies between test and judgmental data?

In its reporting of national performance in the tested subjects (e.g. Ofsted, 2005b) Ofsted simply reports national data without commenting on their validity or reliability (Note 3). They are taken as authoritative and incontestable. In its recent self-evaluation (Matthews and Sammons, 2004) it suggests explanations for its view that ‘overall, school standards – reflected in test results – have risen markedly during the 1990s’, while acknowledging that ‘there has been some levelling off in recent years’ (p.26). It cites:

■ the implementation of the National Literacy and Numeracy Strategies
■ the publication of national comparative performance tables
■ the focus of primary inspections on mathematics and English
■ the ‘perceived distortion of the curriculum that had resulted from the focused application of the strategies in many schools’ (p.29)
■ and in particular improvements in the quality of teaching.

By providing Figure 3, which mirrors the pattern of the test results since 1996, the self-evaluation implies that it is the good or better teaching observed by inspectors that has been a major (or even the major) factor in improving children's performance.

Significantly, it fails to acknowledge fundamental limitations on its inspectors’ ability to make judgements on the quality of teaching, judged in terms of its impact on pupils’ learning. It also does not recognise that at least part of the purported rise in the quality of teaching might have arisen out of inspectors rating highly those lessons involving officially approved teaching methods, irrespective of whether those lessons resulted in enhanced pupil learning (itself very difficult to detect except in unusual circumstances as argued in Richards, 2001). In its list of factors contributing to higher test results it also fails to cite test preparation, ‘teaching to the test’ or technical difficulties with the tests themselves.
6.2.4 QCA test data related to the national strategies

To support Ofsted’s evaluation of the National Literacy and Numeracy Strategies the QCA commissioned the National Foundation for Educational Research (NFER) to conduct yearly testing programmes from 1999 to 2002 (reported in Ofsted, 2002a, 2002b). The programmes focused on pupils in Years 3, 4 and 5 and aimed to provide a detailed picture of changes in standards and progress from Key Stage 1 to Key Stage 2. In all, over 250 primary schools provided data on literacy and a similar number on numeracy.

Figure 3: The proportion of good or better teaching in primary schools

For literacy ‘tests in reading, spelling and writing, similar in content to the QCA optional tests for Years 3, 4 and 5, were specially developed for exclusive use in the evaluation. For reading and spelling, age-standardised scores were calculated’ (Ofsted, 2002a p.37). It is interesting to note that in the executive summary of the technical report no mention is made as to how reflective the test items are of the relevant level descriptions. The performance of whole year groups is compared to look for changes in average age-standardised scores over time. The NFER reports that for spelling and reading there has been consistent gradual improvement in scores, in all three year groups from summer 1999 to summer 2001 – levelling off thereafter. In writing, an upward shift in year 3 and 4 scores has continued until 2002, whilst throughout the period 1999 to 2002 performance in writing has remained steady in Year 5.
For **numeracy** tests in written and mental mathematics, similar in content to the QCA optional tests, were also specially developed for the evaluation. The NFER reports consistent improvements in scores in all three year groups from summer 1999 to summer 2001, levelling off thereafter with the greatest improvement reported at the end of the first year of testing.

The summary of the technical report published in 2002 concludes that for both mathematics and literacy ‘the first three years of the evaluation saw significant improvements in pupils’ performance in all areas’ (Ofsted, 2002a). The meaning accorded to ‘significant’ is not discussed in either statistical or educational terms. Factors that may have contributed to the reported rises in scores are likely to include those already referred to on pages 14 and 15 of this publication (Note 5).

### 6.2.5 Test data from major research projects

Major research projects provide a fourth source of data on standards and performance over time.

**Performance Indicators in Schools Project**

Tymms (2004) reports a set of independent data available from the Performance Indicators in Schools Project (PIPS). He summarises the reading and mathematics attainment scores from the same 122 schools and from over 5,000 pupils each year *using the same test in mathematics and parallel forms of a test in reading* over the period 1997 to 2002. Similar data are also reported for year 4 from 1999 to 2003 using 382 schools and more than 10,000 pupils each year. He states that the tests were specifically written for the PIPS project and are linked to the National Curriculum but as indicated in section 5 it is unclear how far the criteria inherent in such tests relate to the criteria embodied in level descriptions.

The data the tests yielded are more comparable year-on-year than the national tests since in Year 6 the same tests have been used each year in mathematics and in four successive years in reading – instead of different tests each year as in the national testing. The data *might* well be more reliable as an indicator of children’s attainment since the tests used are non-statutory (not ‘high stakes’) and the results therefore less likely to be affected by test preparation. The data for Year 6 from reading test A show only a slight rise from 1997 to 1998 while the reading test data from reading test B show a slight rise from 1999 to 2000 and no change from 2000 to 2002. The reading test data for Year 4 show little change from 2001 to 2004. The Year 6 data for mathematics show a significant rise from 1997 to 2000 levelling off thereafter. The mathematics data for Year 4 show a rise from 2001 to 2003 levelling off for 2004. These results suggest that in Year 6 there was a substantial improvement overall on the PIPS test in mathematics but almost stable performance in reading at a time when the Key Stage 2 results on national tests were rising steadily and between 1997 and 2000 very rapidly.
Leverhulme Numeracy Research Programme

As part of the Leverhulme Numeracy Research Programme data were collected on the performance of children in Year 4. The progress of two cohorts of pupils in 40 schools was monitored. Around 1,300 Year 4 children were tested, one cohort in 1997-98, the other in 2001-02 and their performance compared. Children were tested towards the beginning and end of the school year using the same test for both cohorts.

When the test scores for October and June in Year 4 in 1997-98 are compared with those in 2001-02 there is an average gain in pupils’ results of about three per cent (see Table 2).

Table 2: Mean percentage scores in Leverhulme Year 4 tests 1997-98 and 2001-02

<table>
<thead>
<tr>
<th></th>
<th>October</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997-98</td>
<td>52%</td>
<td>62%</td>
</tr>
<tr>
<td>2001-02</td>
<td>55%</td>
<td>65%</td>
</tr>
<tr>
<td>Rise</td>
<td>3%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Brown *et al.* (2003) comment that ‘this difference is statistically highly significant’ (p.661). This rise is far less than that reported by national testing data for mathematics in Year 6 for the period 1997-02 (see Fig. 2), though it has to be acknowledged that the two sets of tests did not test the same material (and therefore were not working to the same standards) and that two different cohorts were involved with the Year 6 group but not the Year 4 group subject to the pressures of ‘high stakes’ testing. The Year 4 data from the Leverhulme study are more in line with Year 2 national testing data (see Fig. 1) but report a smaller rise in performance. The Leverhulme study’s data more closely correspond to PIPS’ limited data on the performance of children in Year 4 from 2001 to 2004. These contrasts with the results of national testing again raise questions as to the extent of the rise in children’s performance in mathematics from 1997 onwards.

Trends in International Mathematics and Science Survey 2003

Ruddock *et al.* (2004) provide a summary of the 2003 TIMSS survey in which Year 5 and Year 9 pupils in England participated along with their peers from a large number of other countries. Over 3,000 pupils from 123 English schools took part in the 2003 testing programme. A previous cohort of English Year 5 pupils had participated in a similar survey in 1995 and thus it was possible, with limitations, to compare the performance of the two cohorts on those test items common to the tests administered in 1995 and 2003.

On the common mathematical items the scores of the Year 5 pupils are reported as having increased ‘significantly’ (in the statistical sense) from 1995 to 2003 – the rise in the percentage of correct answers being about nine per cent. A similar ‘significant’ increase in
science is reported – with the rise in the percentage of correct answers to common items being in the region of four per cent. How far these common items represent aspects of the National Curriculum level descriptions is not known and therefore it is unclear whether these rises represent an increase in performance against those national standards.

Perhaps too confidently, Ruddock et al. claim that the TIMSS data ‘provide confirmation of improvements in performance in mathematics and science in the late primary years as found in national surveys’ but tellingly they go on to acknowledge that ‘because of the different scales used to report international and national performance levels it is not possible in this report to compare the size of the changes shown in the two arenas’ (p.52) (See note 6).

6.2.6 Other data

Two minor sources of evidence also provide data. Massey et al. (2003) summarise data from six local education authorities (LEAs) where standardised testing data could be linked to children’s achievements in national tests. Massey et al. note that the LEAs used a variety of different standardised tests administered at different stages in the school year, often for different sequences of years. They report that ‘very significantly, there were widespread indications from the various LEAs that standardised test scores have risen between 1996 and 2000. This was true for standardised tests of reading ability and reading comprehension at Key stage 1 and Key stage 2, and, quite possibly, for broadly based standardised mathematics tests at Key stage 2.’ Some what naively, they go on to comment that ‘there seems no reason why teachers should have increased their efforts to prepare children for such tests – and some reasons why they might have taken less interest in doing so over these years’ (p.220) (See Note 7).

A very different picture emerges from standardised test data reported in a small-scale study by Davies and Brember (2001). They claim that there is evidence of stability in mathematics and reading standards in Year 6 over the period 1988 to 1998. However their sample is small and unrepresentative (Year 6 children being tested each year in the same five schools in one LEA) and the tests are not directly related to the National Curriculum level descriptions or programmes of study – being standardised in 1981 in the case of the reading test and in 1986 in the case of the mathematics test.

Overall it is doubtful whether much credence should be placed on the use of such standardised test data in determining whether there has been a rise in performance against National Curriculum standards.
In brief

- the data on performance relate to only three subjects (English, mathematics and science) and only to two age groups (seven-year-olds and eleven-year-olds)
- the national tests show a considerable rise in children’s performance in English and mathematics from 1996 to 2001 followed by a general levelling off thereafter
- this rise in test scores does not necessarily involve a rise in performance against national standards unless these standards have been embodied in the same way and to the same degree in successive tests. However, there is no evidence that this has been the case
- Ofsted has published no inspection evidence on either national standards or performance in relation to those standards. It has simply relied on reporting national test data
- A number of major research projects throw doubt on the considerable rise in performance shown in the national test data.
Overview and policy implications

7.1 Standards and performance: an overview

The question: ‘are standards rising/improving in English primary schools?’ turns out to comprise a number of different questions, none of which can be answered with precision.

The question: ‘are the standards by which children’s performance is assessed rising?’ could be answered authoritatively if a body of ‘expert judgement’ was brought to bear on the level descriptions – the national embodiment of standards or criteria and on the national tests which claim to reflect them. The level descriptions and national tests are public; they have changed over time. Whether those changes represent increases (or decreases) in challenge could be determined but only through the exercise of expert judgment expressed as a consensus view. That expert scrutiny by people whose expertise is credible and whose deliberations are public has not yet taken place.

The question ‘are the standards by which children’s performance are assessed improving? could be answered in a similar fashion. The level descriptions and tests could be scrutinised to see if the changes over time are judged more or less valuable. That expert scrutiny has not yet taken place in a public, accountable way.

The question ‘is children’s performance in relation to national standards rising? could only be answered authoritatively if:

(a) it could be demonstrated publicly that the national standards are embodied in the same way and to the same degree in the tests used – either the same test used on a number of occasions (as, for example, in the PIPS, Leverhulme and TIMSS studies) or in successive tests (as in the national testing programme);

(b) it could be demonstrated that children’s scores have changed significantly (in a statistical sense);

The question ‘is children’s performance in relation to national standards improving?’ could be answered if in addition to (a) and (b) an additional criterion is met i.e:

(c) if it could be judged that these statistically significant changes are likely to be of educational significance.

‘Expert judgement’, almost inevitably expressed in terms of a consensus view (since total agreement is unlikely) is required for (a) and (c); technical expertise in test construction is needed for (a) and (b). So far the requirements in (a) to (c) have not yet been met in a detailed publicly accountable way, either by the QCA or by the research projects cited in this publication.
The studies reviewed in this paper do, however, provide some evidence for changes in performance over time though not in relation to national standards. Because of the different methodologies and tests used the findings of the studies are not strictly comparable.

**However, some tentative intimations are possible:**

In relation to *mathematics*, data from the PIPS project (and from the TIMMS study) suggest a statistically and educationally significant rise in children’s performance in the upper primary years (Years 5 and 6), especially from 1997 to 2000, but levelling off thereafter. However, that rise is not as great as that reported by national testing or by statutory teacher assessments. The data from PIPS and Leverhulme suggest a small rise in performance in Year 4 up to 2002-03 levelling off thereafter. It is not possible to offer any corroboration (or challenge) from research projects to the rise reported for the end of Key Stage 1 based on national testing data.

In relation to *English* the findings are more equivocal. The substantial steady rise from 1996 to 2001 in the Year 6 results on national tests in English is not corroborated by PIPS test data in reading, which show only slight rises from 1997 to 2000. Both sources agree about levelling off in performance thereafter. The reading test data from PIPS for Year 4 show little change from 2001 to 2004. As with mathematics it is not possible to offer any corroboration (or challenge) from research projects to the rise reported for the end of Key Stage 1 based on national testing data.

In relation to *science* there is some evidence of a rise in performance in the upper primary years. TIMSS reports for Year 5 pupils a small but ‘significant’ rise in the percentage of correct answers to common items in tests conducted in 1995 and 2003. This mirrors, although only in part, the substantial rise in national test scores (and to a lesser extent statutory teacher assessments) in Year 6 from 1995 to 2001, followed by a levelling off thereafter. As with English and mathematics it is not possible to offer any corroboration (or challenge) from other research projects to the rise reported for the end of Key Stage 1 based on teacher assessment data.

These very tentative conclusions are consistent with the recently published report from the Statistics Commission, which concludes that ‘it has been established that (a) the improvement in Key stage 2 test scores between 1995 and 2000 substantially overstates improvements in English primary schools over the period, but (b) there nevertheless was some rise in standards’ (2005, p.4).

**Very importantly, there is no published evidence about changes in national standards or performance against those standards in information and communication technology, design technology, history, geography, art, music, physical education or religious education.**
7.2 Policy implications

In the future the nature of national standards and how far they are, or should be, embodied in national tests need to be kept under regular review and procedures established by which performance in relation to those standards can be monitored over time without creating a 'high stakes' backwash effect on schools. This can only be done by harnessing expert judgement, which would be publicly accountable and respected by all relevant stakeholders. That ‘expert’ judgement would need to involve a combination of expertise in the curriculum areas to which the standards apply, in the various phases of education, and in the technology of assessment. It would not be concerned simply with technical matters; it inevitably would involve complex and contentious judgements including value judgements as to what it is valuable or desirable to include in level descriptions (or whatever alternative formulation of national standards is developed in the future). Because of the contentious value-laden nature of the issues it would almost certainly have to operate through consensus, not through unanimity. As Whetton et al (2000) point out standard setting and review rely on society’s and the teaching profession’s willingness to accept the integrity and expertness of those making judgements. There is currently professional (and public) unease on these issues – hence the need for a body representing ‘expert judgement’ that is independent of government. It needs to be established de novo since none of the official agencies, including the QCA and Ofsted, currently enjoys professional confidence in the guardianship of national standards. The micro-politics of how to go about establishing such a body, independent of government but respected and trusted by major stakeholders is, luckily for the author, beyond the scope of this publication.

7.2.1 Keeping standards under review

As long as the current level descriptions provide the basis for national standards the independent body would have to be satisfied that these descriptions are appropriately reflected in the test items used and that this ‘reflection’ is maintained in successive tests until a decision is made to review the level descriptions. Periodically, expert judgement would have to be exercised about the appropriateness of the level descriptions as circumstances change. Judgement would also be required periodically in relation to the level of intellectual challenge provided by the level descriptions, including issues as to whether those levels of challenge are being maintained or whether they need change, either ‘upwards’ or ‘downwards’.

Very importantly too, the independent body should be charged with exploring whether there should be an alternative basis for national standards. If it is convinced of the need for change it should be empowered to propose the necessary steps.
7.2.2 Monitoring performance against standards over time

The independent body would also need to devise a system for assessing performance against those standards – at national level, not at the level of the individual school. It would have to make publicly defensible decisions about which age groups would be involved, which subjects would be included, which aspects of level descriptions would be reflected in tests, and importantly which would not be tested. In order to make comparisons over time both possible and defensible it would need to involve the same tests (split into short sub-tests to avoid overloading individual children) used repeatedly on equivalent samples of pupils of the same age at the same time of the school year. The tests would have to be confidential in order to avoid pressures on schools for test preparation. At any one time only a very small proportion of the school population would need to be tested, with each child being assessed in relation to only part of the level descriptions in one subject. Data at the national level would be published annually. The independent body would also need to consider how to produce a monitoring system designed to be sensitive to the inevitable shifts in curricula and language over time. Tymms (2004) reminds us that the Assessment of Performance Unit established in the 1970s, and ironically disbanded with the onset of national testing, provides a possible model on which to build.

In brief

- The question ‘are standards rising/improving in English primary schools?’ comprises a number of different questions, none of which can be answered with precision
- The evidence suggests that there was some rise in performance in the core subjects between 1995 and 2001, as ‘measured’ by test results, but not as great as national test data (and the government) has suggested
- There is no evidence on how far this change in performance related to national standards (as embodied in level descriptions)
- There is no evidence about changes in children’s performance in the foundation subjects
- An independent body is needed to keep standards under review and to devise a system for assessing performance in relation to those standards over time – at a national level, not at the level of the individual school.
Conclusion

The twin issues of standards and performance in relation to standards are perennial features of educational debate. The proposals discussed in the previous section should help promote a more informed debate in the future but debate there will inevitably be. Standards are too important and too value-laden to be left to the whim of government ministers or self-proclaimed individual 'experts'. It will always be necessary to pose a question similar to ‘are standards rising/improving?’ and to tease out possible answers to the four sub-questions it comprises (section 7.1).

There is, however, a certain peculiarity in the whole standards debate. Whatever the answers teased out, there will always be the same challenges to the teaching profession. Whether standards are rising/improving or falling/deteriorating they could always be made more demanding but should they be? (Note 8) Children's performance in any area of the curriculum could always be improved but should it be? The school system could always do more for the children in its care but in what respects should it? Deliberations about such value-laden issues will always remain unfinished.

Notes

1. For the purposes of this paper ‘performance’ is equated with ‘attainment’ or, to use Ofsted’s term, ‘standards of attainment’.

2. Tellingly, the most recent inspection handbook (Ofsted, 2003) does not define ‘standards’ at all.

3. An example from the 2005 Annual Report illustrates this:

‘Standards in science’

The standards attained in science remained the same as in recent years with 90 per cent of seven-year-olds and 86 per cent of 11-year-olds reaching the expected Levels 2 and 4, respectively. In Key Stage 2 there has been a further two per cent rise in the proportion of pupils achieving Level 5 or above, to 43 per cent. In Key Stage 1 pupils’ attainment is weaker in scientific enquiry than in other attainment targets.

4. Ofsted’s explanation is made in an annex: ‘Inspectors make two separate judgements of standards achieved by pupils:

- standards of work seen – emphasising numeracy and literacy and highlighting strengths and weaknesses in what pupils know, understand and can do;

- how well the pupils achieve, taking account of the progress they have made and other relevant factors.'
The term standards is used to indicate the educational attainment of pupils in relation to clear benchmarks such as National Curriculum levels. When grading the standards of work seen, inspectors judge whether the proportion of pupils achieving the national expectation is below, broadly in line with or above that which is found nationally. This comparison with norms is a key part of the judgement of standards and provides important information for the school being inspected. However, because the inspection grades for attainment are made in comparison with a national norm, when aggregated nationally they can only produce a distribution about that norm, rather than a measure of the national level of attainment of pupils. In this report, evidence from national tests and examinations is used to provide a quantitative measure of the national level of pupils’ attainment’ (Ofsted, 2001 p.89).

5. It is interesting to note that in Ofsted’s latest evaluation report of the national literacy and numeracy strategies (2005b) the summary of the technical report on tests carried out in 2003 and 2004 does not contain an overall summary of changes in pupils’ performance similar to this. The changes it does report are slight.

6. No reference is made in this paper to the 2001 Progress in International Reading Literacy Survey (Mullis et al. 2003) since this does not report changes in performance over time by pupils in English primary schools.

7. However the researchers fail to point out that schools and therefore teachers might have felt themselves under pressure from their LEAs monitoring their performance through the standardised tests and therefore devoted time to test preparation. Also, the test techniques learnt and practised in relation to national and so-called ‘optional’ tests could have helped the children answer the LEA standardised tests.

8. I am grateful to Christopher Schenk for pointing out the crucial distinction between ‘could’ and ‘should’ in these contexts.
References


National Advisory Committee on Creative and Cultural Education. (1999) *All our futures: Creativity, culture and education*. DfEE.


Teaching to Learn Campaign – a great success, a great achievement

What was ATL's Teaching to Learn campaign and what were its aims?

The Teaching to Learn campaign aimed to put learning back at the heart of education policy and shift the debate about schools and schooling away from targets, tests and league tables, and towards learning and the learner.

What did the campaign achieve?

ATL’s 11 specially commissioned reports and several events have continued to provoke debate on the subject. Additionally, government education policy has begun to change in ways which reflect the aims of the campaign:

- teacher assessment at key stage 1 has gained greater importance
- the publication of the Primary National Strategy has begun to focus on the importance of breadth and depth in the curriculum
- the 14-19 White Paper casts doubt on the future of assessment for this age group.

What now for ATL?

With the Government under pressure on assessment, particularly within the ‘Tomlinson agenda’, ATL will continue:

- to call for an end to relentless testing,
- to seek recognition that learning is not just for meeting targets and
- to insist that teachers are given the freedom to enthuse and inspire learners.

Work undertaken during the Teaching to Learn campaign will now lead on to projects which will address the future of the curriculum and accountability across the education system.
“Standards are too important and too value-laden to be left to the whim of government ministers. An independent body is needed to keep standards under review and devise a national system for assessing performance”.

Standards in English Primary Schools: are they rising? by renowned academic Professor Colin Richards concludes that between 1995 and 2001 there was a rise in performance in core subjects as measured by tests results, but that this was not as great as the Government has suggested.

He asks just how valid is the ‘rising standards’ mantra and calls for an independent body to monitor national standards and performance. More fundamentally, Professor Richards shows that the current debate on standards is misleading, and urges the Government to review what it means by standards and how tests reflects those standards.

“This publication is a welcome and long overdue antidote to much of the muddled thinking in which debates about pupil performance are too often couched.”

Dr Mary Bousted, ATL general secretary.