



House of Commons

Children, Schools and Families
Committee

Testing and Assessment

Third Report of Session 2007–08

Volume I



House of Commons

Children, Schools and Families
Committee

Testing and Assessment

Third Report of Session 2007–08

Volume I

Report, together with formal minutes

*Ordered by The House of Commons
to be printed 7 May 2008*

HC 169-I

Published on 13 May 2008
by authority of the House of Commons
London: The Stationery Office Limited
£0.00

The Children, Schools and Families Committee

The Children, Schools and Families Committee is appointed by the House of Commons to examine the expenditure, administration and policy of the Department for Children, Schools and Families and its associated public bodies.

Membership at time Report agreed

Mr Barry Sheerman MP (*Labour, Huddersfield*) (Chairman)
Adam Afriyie MP (*Conservative, Windsor*)
Annette Brooke MP (*Liberal Democrat, Mid Dorset & Poole North*)
Ms Dawn Butler MP (*Labour, Brent South*)
Mr Douglas Carswell MP (*Conservative, Harwich*)
Mr David Chaytor MP (*Labour, Bury North*)
Mr John Heppell MP (*Labour, Nottingham East*)
Mrs Sharon Hodgson MP (*Labour, Gateshead East & Washington West*)
Paul Holmes MP (*Liberal Democrat, Chesterfield*)
Fiona Mactaggart MP (*Labour, Slough*)
Mr Andrew Pelling MP (*Conservative, Croydon Central*)
Mr Andy Slaughter MP (*Labour, Ealing, Acton & Shepherd's Bush*)
Mr Graham Stuart MP (*Conservative, Beverley & Holderness*)
Lynda Waltho MP (*Labour, Stourbridge*)

Powers

The Committee is one of the departmental select committees, the powers of which are set out in House of Commons Standing Orders, principally in SO No 152. These are available on the Internet via www.parliament.uk.

Publications

The Reports and evidence of the Committee are published by The Stationery Office by Order of the House. All publications of the Committee (including press notices) are on the Internet at www.parliament.uk/csf/

Committee staff

The current staff of the Committee are David Lloyd (Clerk), Sarah Thatcher, (Second Clerk), Emma Wisby (Committee Specialist), Judith Boyce (Committee Specialist), Katie Phelan (Committee Assistant), Susan Ramsay (Committee Secretary), and John Kittle (Senior Office Clerk).

Contacts

All correspondence should be addressed to the Clerk of the Children, Schools and Families Committee, House of Commons, 7 Millbank, London SW1P 3JA. The telephone number for general enquiries is 020 7219 6181; the Committee's e-mail address is csfcom@parliament.uk

Contents

Report	<i>Page</i>
Summary	3
Preface	5
1 Introduction: national testing	9
The need for national testing	9
2 The purposes of testing and fitness for purpose	14
Fitness for purpose	14
Validity and reliability	20
Validity	20
Reliability	22
Information for the public	25
3 Targets and performance tables	27
Targets	28
Performance tables	33
The pupil calculation	35
The school calculation	35
4 The consequences of high-stakes uses of testing	41
Teaching to the test	42
Shallow learning	45
Qualifications	46
Is teaching to the test detrimental?	47
Narrowing of the curriculum	49
The burden of testing	52
Pupil stress and demotivation	55
Grade inflation	56
Standards over time	57
Key Stage tests	58
14–19 qualifications	60
International evidence	61
Accountability through sampling	62
5 Reform	67
Single-level tests: the Making Good Progress pilot	67
Personalised learning and Assessment for Learning	70
Progression targets	74
Diplomas	76
The extended project	81
The QCA, development and regulation	82
The Office of the Qualifications and Examinations Regulator	83
Agency for the development of curriculum, assessment and qualifications	83

6 Conclusion	87
Appendix	89
Formal Minutes	98
Witnesses (Volume II)	102
List of written evidence printed (Volume II)	103
List of unprinted written evidence	103
List of Reports from the Committee during the current Parliament	105

Summary

National testing

Systems of national testing have been part of the educational landscape for decades. We consider that a certain amount of national testing at key points in a child's school career is necessary in order to provide a standardised means of measuring educational attainment. However, in recent years the Government has emphasised central control of the education system through testing and associated targets and performance tables, placing test results in a new and more complex context with wide-ranging consequences.

National test results are now used for a wide variety of purposes across many different levels—national, local, institutional and individual. Each of these purposes may be legitimate in its own right, but the question we have asked is whether the current national testing system is a valid means by which to achieve these purposes. We conclude that, in some cases, it is not. In particular, we find that the use of national test results for the purpose of school accountability has resulted in some schools emphasising the maximisation of test results at the expense of a more rounded education for their pupils.

A variety of classroom practices aimed at improving test results has distorted the education of some children, which may leave them unprepared for higher education and employment. We find that 'teaching to the test' and narrowing of the taught curriculum are widespread phenomena in schools, resulting in a disproportionate focus on the 'core' subjects of English, mathematics and science and, in particular, on those aspects of these subjects which are likely to be tested in an examination. Tests, however, can only test a limited range of the skills and activities which are properly part of a rounded education, so that a focus on improving test results compromises teachers' creativity in the classroom and children's access to a balanced curriculum.

The Government's proposals for the new single-level tests may have some positive effects and we approve of the new emphasis on the personalised approach to learning. However, the new regime will continue to use the national tests for the purposes of national monitoring and school accountability. We believe that, without structural modification to address these issues, teaching to the test, narrowing of the taught curriculum and disproportionate focus of resources on pupils on the borderline of targets may continue under the new regime.

Whilst we do not doubt the Government's intention that the National Curriculum should set out "a clear, full and statutory entitlement to learning for all pupils, irrespective of background or ability", we believe that the current system of using a single test for the purposes of measuring pupil attainment, school accountability and national monitoring means that some children receive an education which is focussed too much on those aspects of the curriculum which are subject to national testing.

We conclude that the national testing system should be reformed to decouple these multiple purposes in such a way as to remove from schools the imperative to pursue test results at all costs.

14–19 Diplomas

As the introduction of the new Diplomas approaches, evidence suggests that teachers feel unprepared for the new qualifications and there is anxiety about the limited amount of training they are due to receive. We wonder how schools will collaborate to provide the new curriculum in the competitive environment created by the imperative to show well in performance tables.

Preface

1. Our predecessor Committee, the Education and Skills Committee, announced its call for evidence on testing and assessment on 20 March 2007. That Committee received written evidence over the following months. In the light of that evidence, we decided in November 2007 to continue the inquiry and started taking oral evidence from 10 December 2007.

2. During the inquiry, we took evidence from: Professor Sir Michael Barber, Expert Partner, Global Public Sector Practice, McKinsey and Company; Professor Peter Tymms, Director of Curriculum, Evaluation and Management, School of Education, Durham University; Dr Ken Boston, Chief Executive, Qualifications and Curriculum Authority (QCA); Dr Mary Bousted, General Secretary, Association of Teachers and Lecturers (ATL); Mick Brookes, General Secretary, National Association of Head Teachers (NAHT); Brian Lightman, President, Association of School and College Leaders (ASCL); Keith Bartley, Chief Executive, General Teaching Council for England (GTC); Dr Andrew Bird, Deputy Director General, Assessment and Qualifications Alliance; Murray Butcher, Director, Assessment and Quality, City & Guilds; Jerry Jarvis, Managing Director, Edexcel; Greg Watson, Chief Executive, Oxford, Cambridge and RSA Examinations; Professor Steve Smith, Vice-Chancellor, University of Exeter; Professor Madeleine Atkins, Vice-Chancellor, University of Coventry; Susan Anderson, Director, Human Resources Policy; Richard Wainer, Principal Policy Adviser, Education and Skills, Confederation of British Industry; David Bell, Permanent Secretary; Sue Hackman, Chief Adviser on School Standards; Jon Coles, Director, 14–19 Reform, Department for Children, Schools and Families (DCSF); Jim Knight MP, Minister for Schools and Learners; Ralph Tabberer, Director General, Schools Directorate, Department for Children, Schools and Families.

3. In addition, we received nearly 50 written memoranda, which have helped us with our inquiry. We would like to extend our thanks to our Specialist Advisers, Professor Alan Smithers, Director of the Centre for Education and Employment Research, University of Buckingham, and Professor Dylan Wiliam, Institute of Education, University of London.

4. There are many contested definitions in the field of testing and assessment. For the purposes of this Report we have had to settle on certain definitions, but this is purely for the purposes of conveying our conclusions and such definitions as we have chosen should not be taken to mean that, in a more general sense, we favour any particular school of academic thought over another. Some terms are defined in context elsewhere in this Report. However, it is expedient to define others in advance of the main discussion.

5. We have adopted some of the definitions set out by the QCA in their paper *Evaluating Assessment Systems*. Thus, ‘**assessment**’ is used to refer to “any instrument or process through which student competence or attainment is evaluated (e.g., test, teacher assessment, examination, etc.).”¹ The term ‘**system**’ is used to refer to “the structure and mechanism through which students are assessed”. The system is made up of such details as procedures for test development, distribution, administration, marking, reporting, evaluating, as well as the technical, professional, managerial and administrative employees

required to develop and operate those procedures.² Although not defined by the QCA, we have used the term ‘**testing**’ in this Report to mean a nationally administered assessment, to include Key Stage tests and 14–19 qualifications. It follows that the ‘**testing system**’ refers to the system, as defined by the QCA, which serves the range of national tests.

6. The national testing system in England is complex. The overall structure is summarised in the following two, overlapping tables. Table 1 sets out the regime for the Early Years Foundation Stage and Key Stages 1-4. Table 2 summarises the progression routes for pupils aged 14 to 19.

Table 1: Early Years Foundation Stage and Key Stages 1-4

Age (years)	Year	Stage	Method of assessment at end of Key Stage
3–4	Nursery	Foundation stage	Teacher observation
4–5	Reception class		
5–6	1	Key Stage 1	Primarily teacher assessment. Tests in English and maths in Year 2, marked by teacher.
6–7	2		
7–8	3	Key Stage 2	An element of teacher assessment. Tests in English, maths and science, usually in Year 6, marked by an external marker.
8–9	4		
9–10	5		
10–11	6		
11–12	7	Key Stage 3	An element of teacher assessment. Tests in English, maths and science, usually in Year 9, marked by an external marker.
12–13	8		
13–14	9		
14–15	10	Key Stage 4	GCSEs; other public examinations/qualifications.
15–16	11		

Table 2: 14–19 progression routes

19+	Employment		Higher education		Further education
16-19	Other work-based learning	Advanced Apprenticeships; Apprenticeships	Advanced Diploma Foundation and Higher Diplomas also available	A-levels International Baccalaureate	
Functional skills in all learning routes					
14-16	Young Apprenticeships	Higher Diploma	GCSEs	Foundation learning tier	Foundation Diploma

Source: QCA, *14-19 education and skills: what is a Diploma?*³

7. At the end of Key Stages 1-3, usually at age 7, 11 and 14, children take tests to show what progress they have made throughout that Key Stage. These are known variously as “Key Stage tests”, “National Curriculum tests” or “SATs”, standing for Standard Assessment Tasks. The latter term is still in common usage, although it is strictly anachronistic. Both test results and teacher assessment results are awarded at the end of a Key Stage. Test results are recorded as a series of levels, from 1 to 8. The QCA states that most 7-year-olds will be at level 2; 11-year-olds at level 4; and 14-year-olds at level 5 or 6.

8. These results and the teacher assessments are communicated to parents, along with, for comparison, a summary of test results for children in the child’s age group at their school and the national results for the previous year. The QCA states that Key Stage tests are not “pass or fail” tests, but the level achieved is used to check progress and inform future learning, for example, to indicate where extra help is needed. The results may also be used by some schools for allocating pupils to sets for different subjects. The Government collects the results and uses them as a measure of schools’ performance locally and nationally.

9. At the end of Key Stage 4, children take public examinations, usually GCSEs, and this signals the end of compulsory education. Children who continue in school after the age of 16 may pursue a range of qualifications, including A-levels (now divided into AS and A2 tests spread across two academic years), the International Baccalaureate, an apprenticeship or the new Diplomas.

10. Finally, we set out below a list of common abbreviations used in this report.

AfL	Assessment for Learning
APU	Assessment of Performance Unit
ASCL	Association of School and College Leaders
ATL	Association of Teachers and Lecturers
CVA	Contextual Value Added scores
DCSF	Department for Children, Schools and Families
DfES	Department for Education and Skills
GTC	General Teaching Council for England
IPPR	Institute for Public Policy Research
NAA	National Assessment Agency
NASUWT	National Association of Schoolmasters Union of Women Teachers
NFER	National Foundation for Educational Research
NUT	National Union of Teachers
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessments
QCA	Qualifications and Curriculum Authority
RAISEonline	Reporting and Analysis for Improvement through School self-Evaluation online
TIMSS	Trends in International Mathematics and Science Study

1 Introduction: national testing

The need for national testing

11. Our initial call for evidence for this inquiry asked whether there was a need for a national system of testing in England. In this chapter, we are concerned with the principle of national testing and why it is considered necessary at all. We shall consider in later chapters the purposes for which national tests are used and whether particular instruments of assessment are valid for those purposes.

12. The Chartered Institute of Educational Assessors stated that external testing, in the form of university entrance examinations, was originally used to identify those students who would progress to higher education. As the system evolved, the setting of syllabuses and examinations was delegated to independent examination boards, lightly regulated from the centre. The CIEA continues:

Over the years, as more students stayed in education and took examinations and as competition for places and jobs intensified, the demand for greater comparability across examinations grew and equal access to curriculum and qualifications became the norm. The introduction of National Criteria for GCSE and a National Curriculum and associated assessment arrangements in the 1980s resulted in the centralised system we now have.⁴

13. In 1987, Kenneth Baker, then Secretary of State for Education, announced that there was to be national testing of children at the ages of seven, eleven and fourteen, leading up to the GCSE examinations at sixteen. In his view, the recently announced National Curriculum would be insufficient to improve school standards by itself without measurement of pupils' progress at regular intervals through national testing. The tests were intended to provide objective information to pupils, parents and teachers about what pupils had learned and this, in turn, would enable teachers to identify pupils needing special assistance.⁵ Lord Baker also stated that he wanted test results published and he enshrined this requirement in primary legislation to avoid successor Secretaries of State being "persuaded to go soft" on this aspect of the testing system. He considered that parents wanted access to test result data, but said that publication of results was "anathema to most of the profession", not least because they claimed that test results were not an adequate reflection of the social background of a school, an argument which remains current today. Lord Baker disapproved of this argument, stating that "teachers should not be looking for excuses to explain away poor performance but looking for ways to improve that performance". He insisted on written tests rather than teacher assessments but, when these tests turned out to be "complicated and elaborate", in his words, he was concerned that they would "cause trouble in schools and fail to accomplish [their] objectives". He considered that testing needed to be much simpler.⁶

4 Ev 224

5 Baker, K. (1993). *The Turbulent Years: My Life in Politics*. London: Faber and Faber, pp192 & 199

6 *Ibid.* pp199-200

14. The former Department for Education and Skills told us that, until the introduction of end of Key Stage tests, there were no “objective and consistent performance measures which gave the public confidence about expected standards in primary schools or the intermediary years”.⁷ The Department considered that National Curriculum assessment, together with 16–19 qualifications, provided an “objective and reliable measure of the standards secured by pupils at crucial stages in their development”.⁸ David Bell, the Permanent Secretary at the DCSF, told us in evidence:

I do not accept that we can ever have a system without good and robust national testing and public examinations, the results of which are made available to the public.⁹

15. Ralph Tabberer, Director General of the Schools Directorate at the DCSF, said that the predecessor Department introduced national testing in part because it was felt that the previous system did not provide consistent quality of education across the system. He added that:

I do not know of any teacher or head teacher who would argue against the proposition that education in our schools has got a lot better and more consistent since we introduced national assessment.¹⁰

16. This statement introduces a view that there is a causal link between national testing and apparently rising standards in schools. The DfES stated that:

The benefits brought about by [National Curriculum testing], compared to the time before the accountability of the National Curriculum, have been immense. The aspirations and expectations of pupils and their teachers have been raised. For parents, the benefits have been much better information not only about the progress their own child is making but also about the performance of the school their child attends. And for the education system as a whole, standards of achievement have been put in the spotlight, teachers’ efforts have been directed to make a difference and performance has improved. The public has a right to demand such transparency at a time of record investment in education.¹¹

17. First, then, national testing is considered necessary as a standardised means of validating a pupil’s achievements.¹² Second, accountability, secured through performance indicators derived from the national testing system, is thought to be an important means of driving up standards, leading to confidence in standards amongst users of the education system.

7 Ev 157

8 Ev 157

9 Q327

10 Q395

11 Ev 157

12 See also written evidence from *Heading for Inclusion*, Alliance for Inclusive Education, para 1(g); Ev 226; written evidence from Doug French, University of Hull, para 1.1

18. Others have agreed that national testing contributes to consistency and comparability of testing and assessment across the country, that it allows for monitoring of standards and that it provides a means of assessing the impact of national policy changes on the education system.¹³ At another level, national testing facilitates the development of a shared understanding about learning, which is why a “system-wide approach to the formal recognition and accreditation of learning is a common feature” of many education systems comparable to that in England.¹⁴

19. Associated with the arguments about certifying attainment and accountability is the notion that all pupils should have equal entitlement to a minimum standard of curriculum and associated tests.¹⁵ This means that schools and local authorities should be operating to certain standards of performance in order that children can benefit from the National Curriculum, although it is less clear that national testing is the best or only way to deliver this outcome.¹⁶ Some have agreed with the Government that national testing has been effective in driving up standards.¹⁷ We shall deal with the issue of standards in detail in Chapter 4. However, for the purposes of the present discussion, it is worth noting that many witnesses have taken issue with the idea that national testing is responsible for driving up performance standards in schools.¹⁸

20. School-age national testing can be divided into two categories: National Curriculum testing and qualifications, the latter generally taught and administered in the age range 14–19. The rationales given above for the necessity of national testing generally apply to both categories. There is a further set of rationales which apply to qualifications. It has been argued that a national testing system is needed as a means of certifying a level of achievement for the purposes of higher education and employment.¹⁹ Witnesses have argued that a clear and transparent system of national qualifications is the means to ensure that society can have confidence in the quality and standard of those qualifications.²⁰ Moreover, the NASUWT (National Association of Schoolmasters Union of Women Teachers) argues that a national framework for the accreditation and recognition of learning is essential to ensure the “international transportability of qualifications”. The NASUWT continues:

The work being undertaken by the European Commission on the European Qualifications Framework depends critically on the existence of consistent national examination and assessment systems against which qualifications originating in other countries can be compared. The maintenance of an effective national qualifications system therefore enables learners to access their labour mobility rights

13 Written evidence from Association of Science Education, paras 13-15; written evidence from The Mathematical Association; Ev 246

14 Ev 245

15 Ev 245; Ev 207;

16 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 3(e)

17 Ev 32; Ev 113; Ev 114; Q4

18 See, for example, Ev 53; Ev 262; Ev 198; Ev 247; Q6; written evidence from Association for Science Education, para 22; written evidence from The Mathematical Association

19 Written evidence from Doug French, University of Hull, para 1.1

20 Ev 199; written evidence from The Mathematical Association

as EU citizens and supports the economic and social life of the UK by facilitating the inward migration of qualified workers.²¹

21. So far, the evidence in relation to a system of national testing has been positive. However, some witnesses have pointed out the drawbacks to national testing, including its potential to distort its original purposes. The Association for Science Education has highlighted the danger that:

[...] monitoring of standards leads to enforced compliance in order to meet targets. This in turn results in a culture that limits innovation and enjoyment of learning.²²

Mick Brookes, General Secretary of the National Association of Head Teachers, said that his organisation was in favour of national testing as long as it was “for the right reasons and with the right instruments”, otherwise there was a risk that the curriculum would be distorted.²³

22. These comments aside, we have received little evidence challenging the principle of national testing, although the remainder of this report will be devoted to consideration of the very substantial difficulties of running such a system once it is decided that it is necessary. Whatever the truth about the link between national testing and standards in education, educators accept that accountability of schools is a necessary feature of a modern education system and that national testing has an important part to play.²⁴ Hampshire County Council has said that:

Schools readily acknowledge the need to monitor pupil progress, provide regular information to parents and use assessment information evaluatively for school improvement.²⁵

Mick Brookes told us in evidence:

Nobody in our association wants to return to the 1970s when you did not know what the school up the road was doing, let alone a school at the other end of the country.²⁶

23. We asked the Minister, Jim Knight, about the proposition of a free market in testing and assessment, free of any government involvement or central regulation. He rejected this proposition on the basis that, in his view, there was not a sufficiently large market in testing and assessment for the market to regulate itself effectively. It appears that most countries do, in fact, have some form of centralised, national testing. The QCA provided us with an international comparative analysis of testing and assessment in 20 countries, setting out whether there is a compulsory assessment system; what its purposes are; which pupils are assessed; when they are assessed; and which subjects are assessed. Some key details from a sample of 10 of these countries, including England, are set out in the Appendix. It is

21 Ev 245

22 Written evidence from Association of Science Education, paras 13-15

23 Q128

24 Written evidence from The Mathematical Association; Ev 262; Q129; Q152; Q165

25 Ev 273

26 Q129

interesting to note that England engages exclusively in full-cohort testing, whereas many other jurisdictions make extensive use of sample testing.²⁷

24. In summary, then, the evidence suggests that it is largely uncontroversial that national testing is required for:

- ascertaining and recognising levels of pupil achievement on a standardised basis;
- holding schools and teachers to account;
- assuring the quality of education available to children across the country.

In addition, some have argued that national testing is also needed for:

- promoting confidence in standards;
- providing a basis for parental choice; and
- ascertaining the effects of government policies.

25. We consider that the weight of evidence in favour of the need for a system of national testing is persuasive and we are content that the principle of national testing is sound. Appropriate testing can help to ensure that teachers focus on achievement and often that has meant excellent teaching, which is very welcome.

26. Having accepted the principle of national testing, the remainder of this report will consider some more difficult questions about the structure and operation of the national testing system in England. We will consider the purposes of national testing, what they should be and whether the assessment instruments currently in use are fit for those purposes. We will then discuss performance targets and tables, their uses and consequences for the education system. Amongst these consequences, we identify a number of recurring themes in this inquiry which we discuss in detail, including teaching to the test, narrowing of the taught curriculum and the burden and frequency of testing. Finally, we comment on aspects of some proposals for reform of the testing system: single-level tests, Diplomas and the division of the functions of the QCA.

27 Written evidence from the QCA, Annex 5

2 The purposes of testing and fitness for purpose

27. David Bell, Permanent Secretary at the DCSF, has set out the Department's view of the key purposes of national tests:

We want them to provide objective, reliable information about every child and young person's progress. We want them to enable parents to make reliable and informative judgments about the quality of schools and colleges. We want to use them at the national level, both to assist and identify where to put our support, and also, we use them to identify the state of the system and how things are moving. As part of that, both with national tests and public examinations, we are very alive to the need to have in place robust processes and procedures to ensure standards over time.²⁸

28. The written evidence of the DfES similarly set out a variety of purposes of testing, stating that National Curriculum testing was developed to complement existing public examinations for the 16+ age group and that it is geared towards "securing valid and reliable data about pupil performance, which is used for accountability, planning, resource allocation, policy development and school improvement".²⁹

29. The DfES elaborated on the uses to which data derived from examination results are put. National performance data are used to develop government policy and allocate resources. Local performance data are used for target-setting and to identify "areas of particular under-performance". School performance data form the basis for the findings of inspectors and interventions from School Improvement Partners. Parents make use of school data to make choices about their children's education. The DfES considered that school performance data is an important mechanism for improving school performance and for assisting schools to devise their own improvement strategies. Finally, the DfES stated that examination results for each individual child are "clear and widely-understood measures of progress", which support a personalised approach to teaching and learning and the realisation of each child's potential.³⁰

Fitness for purpose

30. In coming to a view on the government's use of test results for this wide variety of purposes, we have been assisted by the QCA's paper, setting out a framework for evaluating assessment systems.³¹ The QCA highlights the importance of distinguishing the many purposes for which assessment may be used and gives examples of possible interpretations of the term 'purpose of assessment':

28 Q287

29 Ev 157

30 Ev 158-159

31 Ev 21

1. to generate a particular kind of **result**, such as ranking pupils in terms of end-of-course level of attainment;
2. to enable a particular kind of **decision**, such as deciding whether a pupil has learned enough of a particular subject to allow them to move on to the next level;
3. to bring about a particular kind of educational or social **impact**, for example, to compel pupils to learn a subject thoroughly and to compel teachers to align their teaching with the National Curriculum; or the study of GCSE science to support progression to a higher level of study for some pupils and to equip all pupils with sufficient scientific literacy to function adequately as 21st century citizens.³²

31. Clearly, interpretations of the purposes of assessment may be very wide or very narrow, but the important point is that there are a large number of possible purposes. The QCA asks us to consider the uses to which assessment results are put (interpretation 2 above) and distinguishes the four uses set out in the classification scheme established by the Task Group on Assessment and Testing in its 1988 report³³, which are described by the QCA in the following manner:

- formative uses (assessment for learning);
- summative uses (assessment of learning);
- evaluative uses (assessment for accountability); and
- diagnostic uses (assessment for special intervention).³⁴

32. This classification scheme has been used widely in evidence submitted to this inquiry and we, likewise, rely on it extensively in our Report. It should be noted that these categories are not necessarily discreet and the QCA notes many examples of uses to which the results of the national testing system are put which may fall under more than one of the headings of the broad, four-limb classification. The QCA's non-exhaustive list of examples, reproduced at Figure 1, sets out 22 possible uses of assessment results.

32 Ev 23

33 DES (1988), *Task Group on Assessment and Testing: A Report*, London, HMSO

34 Ev 23

Figure 1 Some examples of the uses to which assessment results can be put³⁵

1. **social evaluation** (to judge the social or personal value of students' achievements)
2. **formative** (to identify students' proximal learning needs, guiding subsequent teaching)
3. **student monitoring** (to decide whether students are making sufficient progress in attainment in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
4. **diagnosis** (to clarify the type and extent of students' learning difficulties in light of well-established criteria, for intervention)
5. **provision eligibility** (to determine whether students meet eligibility criteria for special educational provision)
6. **screening** (to identify students who differ significantly from their peers, for further assessment)
7. **segregation** (to segregate students into homogeneous groups, on the basis of aptitudes or attainments, to make the instructional process more straightforward)
8. **guidance** (to identify the most suitable courses, or vocations for students to pursue, given their aptitudes)
9. **transfer** (to identify the general educational needs of students who transfer to new schools)
10. **placement** (to locate students with respect to their position in a specified learning sequence, to identify the level of course which most closely reflects it)
11. **qualification** (to decide whether students are sufficiently qualified for a job, course or role in life—that is, whether they are equipped to succeed in it—and whether to enrol them or to appoint them to it)
12. **selection** (to predict which students—all of whom might, in principle, be sufficiently qualified—will be the most successful in a job, course or role in life, and to select between them)
13. **licensing** (to provide legal evidence—the licence—of minimum competence to practice a specialist activity, to warrant stakeholder trust in the practitioner)
14. **certification** (to provide evidence—the certificate—of higher competence to practise a specialist activity, or subset thereof, to warrant stakeholder trust in the practitioner)
15. **school choice** (to identify the most desirable school for a child to attend)
16. **institution monitoring** (to decide whether institutional performance—relating to individual teachers, classes or schools—is rising or falling in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
17. **resource allocation** (to identify institutional needs and, consequently, to allocate resources)
18. **organisational intervention** (to identify institutional failure and, consequently, to justify intervention)
19. **programme evaluation** (to evaluate the success of educational programmes or initiatives, nationally or locally)
20. **system monitoring** (to decide whether system performance—relating to individual regions or the nation—is rising or falling in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
21. **comparability** (to guide decisions on comparability of examination standards for later assessments on the basis of cohort performance in earlier ones)
22. **national accounting** (to 'quality adjust' education output indicators)

Source: QCA.

33. Each one of these possible uses of assessment results can, in itself, be seen as a purpose of assessment, depending on the context. Where an assessment instrument is designed and used only for one purpose, the answer to the question “is it fit for purpose” is the result of a relatively straightforward process of evaluation. However, the government’s evidence, set out in paragraphs 27-29 above, highlights the fact that national tests are used for a wide variety of purposes at a number of different levels: national, local, school and individual.

34. Each instrument of assessment is (or should be) designed for a specific purpose or related purposes. It will only be fit (or as fit as a test instrument can be) for those purposes for which it is designed. The instrument will not necessarily be fit for any other purposes for which it may be used and, if it is relied upon for these other purposes, then this should be done in the knowledge that the inferences and conclusions drawn may be less justified than inferences and conclusions drawn from an assessment instrument specifically designed for those purposes.³⁶

35. The DfES recognised that an assessment system inevitably makes trade-offs between purposes, validity, reliability and manageability. However, the evidence from the DfES and the DCSF has been consistent: that the data derived from the current testing system “equips us with the best data possible to support our education system”.³⁷ David Bell, Permanent Secretary at the DCSF, told us that:

I think that our tests give a good measure of attainment and the progress that children or young people have made to get to a particular point. It does not seem to be incompatible with that to then aggregate up the performance levels to give a picture of how well the school is doing. Parents can use that information, and it does not seem to be too difficult to say that, on the basis of those school-level results, we get a picture of what is happening across the country as a whole. While I hear the argument that is often put about multiple purposes of testing and assessment, I do not think that it is problematic to expect tests and assessments to do different things.³⁸

36. Dr Ken Boston of the QCA told us that the current Key Stage tests were fit for the purpose for which they were designed, that is, “for cohort testing in reading, writing, maths and science for our children at two points in their careers and for reporting on the levels of achievement”.³⁹ The primary purpose of Key Stage tests was “to decide the level that a child has reached at the end of a Key Stage”.⁴⁰ He explained that Key Stage tests are developed over two and a quarter years, that they are pre-tested and run through teacher panels twice and that the marking scheme is developed over a period of time. He considers that Key Stage tests are as good as they can be and entirely fit for their design purpose.⁴¹ Dr Boston noted, however, that issues were raised when, having achieved a test which is fit for

36 Ev 24-25

37 Ev 157

38 Q290

39 Q79

40 Q84

41 Q79; see also Ev 31

one purpose, it is then used for other purposes. Figure 1 above lists 22 purposes currently served by assessments and, of those, 14 are being served by Key Stage tests.

My judgment is that, given that there are so many legitimate purposes of testing, and [Figure 1 above] lists 22, it would be absurd to have 22 different sorts of tests in our schools. However, one serving 14 purposes is stretching it too far. Three or four serving three or four purposes each might get the tests closer to what they were designed to do. [...] when you put all of these functions on one test, there is the risk that you do not perform any of those functions as perfectly as you might. What we need to do is not to batten on a whole lot of functions to a test, but restrict it to three or four prime functions that we believe are capable of delivering well.⁴²

37. Similarly, Hargreaves et al argue that one test instrument cannot serve all the Government's stated purposes of testing because they conflict to a certain extent, so that some must be prioritised over others. According to them, the purpose of assessment for learning has suffered at the expense of the other stated purposes whereas, in their view, it should have priority.⁴³ The conflicts between the different purposes are not, perhaps, inherent, but arise because of the manner in which people change their behaviour when high-stakes are attached to the outcomes of the tests. Many others have raised similar points, claiming that two purposes in particular, school accountability on the one hand and promoting learning and pupil progress on the other, are often incompatible within the present testing system.⁴⁴ The practical effects of this phenomenon will be discussed further in Chapter 4. However, we have been struck by the depth of feeling on this subject, particularly from teachers.

38. The GTC (General Teaching Council for England) argues that reliance on a single assessment instrument for too many purposes compromises the reliability and validity of the information obtained. It claims that the testing system creates tensions that "have had a negative impact upon the nature and quality of the education" received by some pupils. It concludes that "These tensions may impede the full realisation of new approaches to education, including more personalised learning".⁴⁵

39. The NUT (National Union of Teachers) stated that successive governments have ignored the teaching profession's concerns about the impact of National Curriculum testing on teaching and learning and it believes that this is "an indictment of Government attitudes to teachers' professional judgment".⁴⁶ The NUT argues further that:

It is the steadfast refusal of the Government to engage with the evidence internationally about the impact of the use of summative test results for institutional evaluation which is so infuriating to the teaching profession.⁴⁷

42 Q79

43 "System Redesign—2: assessment redesign", David Hargreaves, Chris Gerry and Tim Oates, November 2007, pp28-29

44 Ev 261; Ev 264; Ev 198; Ev 273; Ev 75; Ev 47; Q134; Q237; written evidence from Association of Science Education, paras 5-6; written evidence from The Mathematical Association, under headings "General Issues" and "National Key Stage Tests";

45 Ev 75

46 Ev 261

47 Ev 264

40. An NUT study, published in 2003, found that the use of test results for the purpose of school accountability had damaging effects on teachers and pupils alike. Teachers felt that the effect was to narrow the curriculum and distort the education experience of pupils. They thought that the “excessive time, workload and stress for children was not justified by the accuracy of the test results on individuals”⁴⁸.

41. Others have argued that the use of national testing for the twin aims of pupil learning and school accountability has had damaging effects on children’s education experience. Hampshire County Council accepts that tests are valuable in ascertaining pupil achievement but is concerned that their increasingly extensive use for the purposes of accountability “has now become a distraction for teachers, headteachers and governing bodies in their core purpose of educating pupils”.⁴⁹ The Council continues:

Schools readily acknowledge the need to monitor pupil progress, provide regular information to parents and use assessment information evaluatively for school improvement. The key issue now is how to balance the need for accountability with the urgent need to develop a fairer and more humane assessment system that *genuinely* supports good learning and teaching.⁵⁰

42. It is not a necessary corollary of national testing that schools should narrow the curriculum or allow the tests to dominate the learning experience of children, yet despite evidence that this does not happen in all schools there was very wide concern that it is common. We return to these concerns in Chapter 4.

43. The NUT highlighted evidence which suggests that teachers feel strongly that test results do not accurately reflect the achievements of either pupils or a school.⁵¹ The NAHT considers that Key Stage tests provide one source of helpful performance data for both students and teachers, but that it is hazardous to draw too many conclusions from those data alone. They argue that “A teacher’s professional knowledge of the pupil is vital—statistics are no substitute for professional judgment”.⁵² On the subject of school performance, the NAHT states that Key Stage test results represent only one measure of performance amongst a wide range, from financial benchmarking through to full Ofsted inspections. It considers that self-evaluation, taken with other professional educational data, “is far more reliable than the one-dimensional picture which is offered by the SATs”.⁵³ The Association of Colleges stated that performance tables constructed from examination results data do not adequately reflect the actual work of a school and that the emphasis on performance tables risks shifting the focus of schools from individual need towards performance table results.⁵⁴

48 Ev 263

49 Ev 272

50 Ev 273

51 Ev 263

52 Ev 68

53 Ev 69

54 Ev 198

44. The evidence we have received strongly favours the view that national tests do not serve all of the purposes for which they are, in fact used. The fact that the results of these tests are used for so many purposes, with high-stakes attached to the outcomes, creates tensions in the system leading to undesirable consequences, including distortion of the education experience of many children. In addition, the data derived from the testing system do not necessarily provide an accurate or complete picture of the performance of schools and teachers, yet they are relied upon by the Government, the QCA and Ofsted to make important decisions affecting the education system in general and individual schools, teachers and pupils in particular. In short, we consider that the current national testing system is being applied to serve too many purposes.

Validity and reliability

45. If the testing system is to be fit for purpose, it must also be valid and reliable.⁵⁵ City and Guilds, an Awarding Body accredited by the QCA, has told us:

[...]there is considerable obligation on the designer of tests or assessments to make them as efficient and meaningful as possible. Assessment opportunities should be seen as rare events during which the assessment tool must be finely tuned, accurate and incisive. To conduct a test that is inaccurate, excessive, unreliable or inappropriate is unpardonable.⁵⁶

46. Although there is no consensus in the evidence on the precise meanings of the terms ‘validity’ and ‘reliability’, we have had to come to a working definition for our own purposes. ‘**Validity**’ is at the heart of this inquiry and we take it to refer to an overall judgment of the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. This judgment is based on the premise that the tests in fact measure what it is claimed that they measure or, as the NFER (National Foundation for Educational Research) puts it, “the validation of a test consists of a systematic investigation of the claims being made for it”.⁵⁷

47. Our definition of validity is a broad definition precisely because it includes the concept of reliability: an assessment system cannot be valid without being reliable. ‘**Reliability**’ we define as the ability to produce the same outcome for learners who reach the same level of performance.

Validity

48. If a valid test is defined as one that actually measures what it is claimed it measures, the NFER considered that Key Stage tests would be valid if they “give an accurate and useful indication of students’ English, science or mathematical attainment in terms of National Curriculum levels”.⁵⁸ The NFER made the following assessment of the coverage of the total curriculum:

55 Ev 233

56 Ev 110-111

57 Ev 257

58 Ev 257

The tests do have limited coverage of the total curriculum: the English tests omit Speaking and Listening, the science tests formally omit the attainment target dealing with scientific enquiry (though questions utilising aspects of this are included) and mathematics formally omits using and applying mathematics. Outside of these the coverage of content is good. The fact that the tests change each year means that the content is varied and differing aspects occur each year.⁵⁹

The NFER stated that the current tests adequately serve the accountability purposes of testing. They may not meet so successfully the standards of validity necessary for the purpose of national monitoring, although the NFER believed that the tests are as good as they can be for this purpose. The NFER said that, in principle, if there was to be an assessment system with the sole purpose of national monitoring of standards using comparable measures, then a low-stakes, lightly-sampled survey was probably the most valid form of assessment.

49. The validity of the current testing system has elsewhere been repeatedly challenged in the evidence to this inquiry. Whilst asserting that the Key Stage tests are fit for purpose, the QCA has acknowledged that:

Like any tests, however well designed, they can measure only a relatively narrow range of achievement in certain subjects on a single occasion and they cannot adequately cover some key aspects of learning.⁶⁰

50. Many witnesses are less content than the NFER with coverage of the National Curriculum and have challenged the validity of national tests on grounds that they test only a narrow part of the set curriculum and a narrow range of a pupil's wider skills and achievements.⁶¹ It is also argued that existing tests measure recall rather than knowledge⁶² and neglect skills which cannot easily be examined by means of an externally-marked, written assessment.⁶³ Furthermore, to enhance the ability of pupils to recall relevant knowledge in an examination, thereby improving test scores, teachers resort to coaching, or 'teaching to the test',⁶⁴ and to teaching only that part of the curriculum which is likely to be tested in an examination.⁶⁵ The Government does not intend it, but it is undeniable that the high stakes associated with achieving test benchmarks has led schools and teachers to deploy inappropriate methods to maximise the achievement of benchmarks. This is examined in Chapter 4. For now, we note that these phenomena affect the validity of the examination system as a whole, not just test instruments in particular, because the education experience of a child is arguably directly affected by the desire of some teachers and schools to enhance their pupils' test results at the expense of a more rounded education.

59 Ev 257-258

60 Ev 32

61 Ev 56; Ev 71; Q128; written evidence from the Advisory Committee on Mathematics Education, paras 18-20; written evidence from Association for Achievement and Improvement through Assessment, para 4

62 Ev 263; Ev 269; written evidence from Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hants

63 Ev 238; Ev 239; written evidence from Doug French, University of Hull, para 2.2; written evidence from Association for Achievement and Improvement through Assessment, para 4

64 Ev 75; Ev 56; written evidence from Doug French, University of Hull, para 1.3

65 Ev 60; Ev 75; Ev 232; Q139; Written evidence from Doug French, University of Hull, para 1.3

Reliability

51. Professors Black, Gardner and Wiliam argued that the reliability of national tests and testing systems is limited; that the results of such systems are misused; and that the effects of such misuse would be reduced if test developers were required to inform the public of the margins of error inherent in these testing systems. They stressed that limited reliability of testing systems is systemic and inevitable and does not imply lack of competence or professionalism on the part of test developers.⁶⁶ The results of any assessment system are subject to measurement error because they are based on a limited sample of a candidate's attainment. In order that the testing system should be manageable and affordable, only a limited number of questions can be set, to be answered in a limited time and on a given day. Variations in results for a given candidate will arise out of the particular topics and skills tested in the particular test instrument and out of the performance of the candidate on the day. Other evidence has suggested that children aged ten or eleven exhibit increased tension and stress when facing a week of examinations in which they are expected to demonstrate "the full extent of their learning from seven years of education".⁶⁷ This may affect examination performance. Black et al stated that the 'true score' of a candidate can never be known because it is practically impossible to test more than a limited sample of his or her abilities.⁶⁸ Indeed, their evidence was that up to 30% of candidates in any public examination in the UK will receive the wrong level or grade, a statistical estimate which has also been quoted by others in evidence.⁶⁹ Dr Boston of the QCA accepted that error in the system exists, but said he was surprised by a figure as high as 30%.⁷⁰ Jon Coles, Director of 14–19 Reform at the DCSF, told us that:

[...] I simply do not accept that there is anything approaching that degree of error in the grading of qualifications, such as GCSEs and A-levels. The OECD has examined the matter at some length and has concluded that we have the most carefully and appropriately regulated exam system in the world.⁷¹

[...] I can say to you without a shadow of a doubt—I am absolutely convinced—that there is nothing like a 30% error rate in GCSEs and A-levels.⁷²

52. We suspect that the strength of this denial stemmed from a misunderstanding of the argument made by Black et al. In their argument, they make the assumptions that tests are competently developed and that marking errors are minimal.⁷³ The inherent unreliability of the tests stems from the limited knowledge and skills tested by the assessment instrument and variations in individuals' performance on the day of the test.⁷⁴ This does not impugn the work of the regulator or the test development agencies and very little can

66 Ev 202-203

67 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 2(a)

68 Ev 203

69 Ev 61; Ev 75; Ev 221-222; Ev 226; Q128

70 Q83

71 Q297

72 Q298

73 Ev 202-203

74 Ev 203

be done to enhance reliability whilst maintaining a manageable and affordable system. The NFER gave similar evidence that the current Key Stage tests:

[...] have good to high levels of internal consistency (a measure of reliability) and parallel form reliability (the correlation between two tests). Some aspects are less reliable, such as the marking of writing, where there are many appeals/reviews. However, even here the levels of marker reliability are as high as those achieved in any other written tests where extended writing is judged by human (or computer) grades. The reliability of the writing tests could be increased but only by reducing their validity. This type of trade off is common in assessment systems with validity, reliability and manageability all in tension.⁷⁵

53. Black et al identify that reliability could theoretically be enhanced in a number of ways:

- Narrowing the range of question types, topics and skills tested; but the result would be less valid and misleading in the sense that users of that information would have only a very limited estimate of the candidates' attainments.
- Increasing the testing time to augment the sample of topics and skills tested; however, reliability increases only marginally with test length.⁷⁶ For example, to reduce the proportion of pupils wrongly classified in a Key Stage 2 test to within 10%, it is estimated that 30 hours of testing would be required. (The NFER expressed the view that the present tests provide as reliable a measurement of individuals as is possible in a limited amount of testing time.⁷⁷)
- Collating and using information that teachers have about their pupils. Teachers have evidence of performance on a range of tasks, in many different topics and skills and on many different occasions.

54. Black et al conclude this part of their argument by stating that, when results for a group of pupils are aggregated, the result for the group will be closer to the 'true score' because random errors for individuals—which may result in either higher or lower scores than their individual 'true score'—will average out to a certain extent.⁷⁸ The NFER went further, stating that aggregated results over large groups such as reasonably large classes and schools give an "extremely high" level of reliability at the school level.⁷⁹ Nevertheless, Black et al argue that not enough is known about the margins of error in the national testing system. Professor Black wrote to the QCA to enquire whether there was any research on reliability of the tests which it develops:

The reply was that "there is little research into this aspect of the examining process", and [the QCA] drew attention only to the use of borderline reviews and to the reviews arising from the appeals system. We cannot see how these procedures can be of defensible scope if the range of the probable error is not known, and the evidence

75 Ev 257

76 See also Ev 236

77 Ev 257

78 Ev 203-204; see also Ev 226

79 Ev 257

suggests that if it were known the volume of reviews needed would be insupportable.⁸⁰

55. Black et al go on to argue that it is profoundly unsatisfactory that a measure of the error inherent in our testing system is not available, since important decisions are made on the basis of test results, decisions which will be ill-judged if it is assumed that these measures are without error. In particular, they argue that current policy is based on the idea that test results are reliable and teachers' assessments are unreliable. They consider that reliability could, in fact, be considerably enhanced by combining the two effectively and that work leading in this direction should be prioritised.⁸¹ Black et al conclude that:

[...] the above is not an argument against the use of formal tests. It is an argument that they should be used with understanding of their limitations, an understanding which would both inform their appropriate role in an overall policy for assessment, and which would ensure that those using the results may do so with well-informed judgement.⁸²

56. Some witnesses have emphasised what they see as a tension between validity and consistency in results. The argument is that, over time, national tests have been narrowed in scope and marking schemes specified in an extremely detailed manner in order to maximise the consistency of the tests. In other words, candidates displaying the same level of achievement in the test are more likely to be awarded the same grade since there is less room for the discretion of the examiner. However, it is argued further that this comes at the expense of validity, in the sense that the scope of the tests are narrowed so much that they test very little of either the curriculum or the candidate's wider skills.⁸³ Sue Hackman, Chief Adviser on School Standards at the DCSF, recognised this trade-off. However, she also told us that in relation to Key Stage tests the Department, together with the QCA, has tried to include a range of questions in test papers, some very narrow and others rather wider. In this way, she considered that a compromise has been reached between "atomistic and reliable questions, and wide questions that allow pupils with flair and ability to show what they can do more widely".⁸⁴

57. Many witnesses have called for greater emphasis on teacher assessment in order to enhance both the validity and the reliability of the testing system.⁸⁵ A move towards a better balance between regular, formative teacher assessment and summative assessments—the latter drawn from a national bank of tests, to be externally moderated—would provide a more rounded view of children's achievements, and many have criticised the reliance on a 'snapshot' examination at a single point in time.⁸⁶

80 Ev 204

81 Ev 204-205

82 Ev 205

83 Ev 226

84 Q324

85 Ev 112; Ev 204; Ev 205; Ev 223; Ev 239; Ev 271; written evidence from the Association for Achievement and Improvement through Assessment, para 5

86 Ev 49; Ev 68; Ev 75; Ev 112; Ev 223; Ev 225; Ev 271; written evidence from Purbrook Junior School, Waterloo, para 5; written evidence from Association for Achievement and Improvement through Assessment, paras 4-5

58. We consider that the over-emphasis on the importance of national tests, which address only a limited part of the National Curriculum and a limited range of children's skills and knowledge has resulted in teachers narrowing their focus. Teachers who feel compelled to focus on that part of the curriculum which is likely to be tested may feel less able to use the full range of their creative abilities in the classroom and find it more difficult to explore the curriculum in an interesting and motivational way. We are concerned that the professional abilities of teachers are, therefore, under-used and that some children may suffer as a result of a limited educational diet focussed on testing. We feel that teacher assessment should form a significant part of a national assessment regime. As the Chartered Institute of Educational Assessors states, "A system of external testing alone is not ideal and government's recent policy initiatives in progress checks and diplomas have made some move towards addressing an imbalance between external testing and internal judgements made by those closest to the students, i.e. the teachers, in line with other European countries".⁸⁷

Information for the public

59. The National Foundation for Educational Research stated that no changes should be made to the national testing system without a clear statement of the purposes of that system in order or priority. The level of requirements for validity and reliability should be elucidated and it should be made clear how these requirements would be balanced against the need for manageability and cost-effectiveness.⁸⁸ The NFER commented that Key Stage testing in particular:

[...] is now a complex system, which has developed many different purposes over the years and now meets each to a greater or lesser extent. It is a tenet of current government policy that accountability is a necessary part of publicly provided systems. We accept that accountability must be available within the education system and that the assessment system should provide it. However, the levels of accountability and the information to be provided are open to considerable variation of opinion. It is often the view taken of these issues which determines the nature of the assessment system advocated, rather than the technical quality of the assessments themselves.⁸⁹

60. Cambridge Assessment criticised agencies, departments and Government for exaggerating the technical rigour of national assessment. It continued:

[...] any attempts to more accurately describe its technical character run the risk of undermining both the departments and ministers; '[...] if you're saying this now, how is it that you said that, two years ago [...]'. This prevents rational debate of problems and scientifically-founded development of arrangements.⁹⁰

87 Ev 222

88 Ev 251

89 Ev 251

90 Ev 251

Cambridge Assessment stated further that international best practice dictates that information on the measurement error intrinsic to any testing system should be published alongside test data and argues that this best practice should be adopted by the Government.⁹¹ Professor Peter Tymms of Durham University similarly argued that:

[...] it would certainly be worth trying providing more information. I think that the Royal Statistical Society's recommendation not to give out numbers unless we include the uncertainties around them is a very proper thing to do, but it is probably a bit late.⁹²

61. We are concerned about the Government's stance on the merits of the current testing system. We remain unconvinced by the Government's assumption that one set of national tests can serve a range of purposes at the national, local, institutional and individual levels. We recommend that the Government sets out clearly the purposes of national testing in order of priority and, for each purpose, gives an accurate assessment of the fitness of the relevant test instrument for that purpose, taking into account the issues of validity and reliability.

62. We recommend further that estimates of statistical measurement error be published alongside test data and statistics derived from those data to allow users of that information to interpret it in a more informed manner. We urge the Government to consider further the evidence of Dr Ken Boston, that multiple test instruments, each serving fewer purposes, would be a more valid approach to national testing.

91 Ev 214-215

92 Q19

3 Targets and performance tables

63. In the consultation document *Making Good Progress*, the Government refers to “the framework of tests, targets and performance tables which have helped drive up standards so sharply over the past decade”.⁹³ Key Stage tests are used to generate data on pupil performance at the end of each Key Stage. This data is then collated and used to “measure trends across time, across schools, and by almost every conceivable characteristic of the pupils”. The Government’s main focus at present is on improvement measured by average achievement across a school at the end of each Key Stage and, to this end, schools are given targets.⁹⁴ The results for each school are aggregated into Achievement and Attainment Tables which allow comparison of school results, either in terms of absolute test scores or in terms of Contextual Value Added scores (a measure of school performance adjusted for a variety of factors which have been statistically shown to have an impact on an individual’s progress: see further paragraphs 90-103 below). These tables are also referred to as “performance tables” (the term we have adopted as a shorthand) and “league tables”. The latter term is, perhaps, less accurate because the tables published by the DCSF do not rank schools in order of achievement, although ranked tables are compiled and published by the news media annually.⁹⁵

64. The Government states that schools are provided with a number of tools to understand and track attainment:

- the DCSF provides tools to track the attainment of individual pupils adjusted for their starting point (value added) and for social factors (Contextual Value Added);
- Local Authorities provide Fischer Family Trust estimates of future pupil attainment, based on prior attainment;
- the DCSF provides schools with the Autumn Package, an annual set of data, together with an analytic tool, the Pupil Attainment Tracker, to assist in measuring the progress of individual pupils;
- these materials can now be accessed online through the RAISEonline system (Reporting and Analysis for Improvement through School self-Evaluation).⁹⁶

65. Witnesses to this inquiry have, however, challenged the Government’s assertions that tests, targets and performance tables have driven up standards. The NASUWT said that:

There is little evidence that performance tables have contributed to raising standards of attainment. A growing number of international studies show that other comparable education systems, including those in Wales, Scotland and Northern

93 (2006) Department for Education and Skills, *Making Good Progress: How can we help every pupil to make good progress at school?*, p2.

94 *ibid.* p4

95 For examples, see http://www.timesonline.co.uk/tol/life_and_style/education/a_level_gcse_results/ and <http://www.telegraph.co.uk/education/main.jhtml?xml=/education/leaguetables/primary2006/resultsall.xml>

96 *ibid.* p4

Ireland, have reached and maintained high educational standards without use of the performance tables.⁹⁷

66. The NUT argued that there is no evidence that they have had such an effect. The ATL (Association of Teachers and Lecturers) also noted the Government's assertions of improving standards, but questioned "whether this means that our pupils are learning more and better". It referred to research at Durham University suggesting that pupils who reach Level 4 at Key Stage 2 do not retain what they have learned over a period of six months to a year.⁹⁸ The ASCL (Association of School and College Leaders) considers that the aggregation of individual test scores creates a high-stakes testing system which will inevitably create a false picture of progress.⁹⁹ The ASCL argues further that the Government has produced no evidence to support the contention that targets and performance tables have driven up standards in recent years, a contention "which has taken on the aspect of a dogma".¹⁰⁰ The ASCL considers it more likely that increased investment in the education system, leading to better leadership, staffing and facilities, has led to improved performance.¹⁰¹

Targets

67. Targets for pupil attainment are set with the intention that children should meet the expected levels for their age in the core subjects of English, mathematics and science. Ralph Tabberer, Director General of the Schools Directorate at the DCSF, emphasised that targets are set at a level which ensures that the pupil achieving it is ready to move on to the next stage of schooling. For example, Level 4 is set as the target at the end of Key Stage 2 because that is the level which ensures the pupil is ready to move on to the secondary curriculum.¹⁰²

68. The IPPR (Institute for Public Policy Research) notes the considerable emphasis the Government has placed on what the IPPR terms the 'standards agenda', that is, increasing the proportion of pupils achieving target levels in Key Stage tests and on minimising the number of schools who do not meet the targets. The Government has placed its faith in a "quasi-market in school places", in which parental choice should drive up standards, with targets and performance tables placed at the heart of this mechanism. Value added measures and CVA are intended to provide more context for the raw results.¹⁰³ The IPPR continues:

Results are now used to inform school decisions about performance-related pay, to inform Ofsted decisions about whether schools should be given light or heavy touch inspections and, combined with targets, to inform judgments about the efficacy of educational initiatives such as the Primary Strategies.¹⁰⁴

97 Ev 247

98 Ev 60-61

99 Ev 48

100 Ev 53

101 Ev 53

102 Q370

103 Ev 234

104 Ev 234

69. Further, the IPPR notes that test data is now aggregated into sophisticated data banks, most recently RAISEonline, allowing for performance monitoring, progress evaluation, problem diagnosis, resource allocation and target setting based on “a full understanding of what might be achievable”¹⁰⁵. This, it is argued, has facilitated top-down performance management by government, but also allows local authorities and schools to set attainment targets for their pupils, to assess pupil progress against that of similar pupils elsewhere and to assess school results against national comparators. It is even possible to make comparisons at the level of individual test questions. The IPPR considers that this is a powerful tool for “supporting informed and rigorous self-management”.¹⁰⁶

70. Targets, according to the Government, are a primary means of focussing the efforts of schools and teachers on achieving these ends. *Making Good Progress* sets out clearly the Government’s aims for the education system and the relevance of targets:

The reason for pursuing higher standards is not in order to achieve numerical targets or to deliver accountability. Useful and necessary as these are, they are the servants and not the masters. The data and targets we set are the means towards the objective of equipping pupils with the skills and knowledge they need: education for self-fulfilment, access and equality. So it is important that we use our data and set our accountability targets to achieve the ends we most value.¹⁰⁷

The Minister, Jim Knight, told us that schools take their performance in tests very seriously and that this drives the priorities of English, mathematics and science. He stated that there was evidence that sharp accountability has contributed to improvement in those areas.¹⁰⁸ The DfES stated that:

The publication of threshold measures of performance is a strong incentive for schools and colleges to ensure that as many pupils/students as possible achieve the required standard, particularly at Key Stages 1-3, in the core subjects of English, mathematics and science.¹⁰⁹ [...]

Used together, threshold and CVA measures provide a powerful tool for school improvement and raising standards across the education system, enabling us to track changes in performance over time nationally and locally, and at school and individual pupil level [...]¹¹⁰

71. The Minister told us that targets have been put in place so that 30% of pupils are expected to achieve five GCSEs at grades A* to C, including English and mathematics. The Minister said that, a decade ago, half of all schools did not have more than 30% of pupils

105 Ev 235

106 Ev 235

107 *ibid.* p4

108 Q367

109 Ev 158

110 Ev 158-159

achieving five GCSEs at grades A* to C, but that figure is now down to 21% of schools. This measure is used to target schools needing improvement.¹¹¹

72. Much of the criticism directed at national tests actually derives from the distortions created by performance targets. When the current administration assumed responsibility for the 'delivery' of education, the accountability structures which were put in place were based on pupil performance in national tests. Pressure was applied to the system by means of targets and performance tables, with educational outputs regularly measured and a wide variety of strategies and initiatives put in place to increase productivity.¹¹² Viewed in this light, targets based on national, summative test results are not the servant but are the engine which drives productivity in the education system.

73. Test results are not the output of education, but a proxy for the education taking place every day in classrooms across the country. OCR, one of the Awarding Bodies accredited by the QCA, argues that problems arise when test results, designed to measure pupil attainment, are used as a proxy measure for different purposes:

The use of qualifications in school performance tables, national targets, OECD comparisons etc leads to misinformation and drives undesirable behaviours.¹¹³

74. When high-stakes are attached to the proxy, rather than the education it is meant to stand for, distortion may occur in the shape of teaching to the test, narrowing of the taught curriculum to those subjects likely to be the subject of examination and an inappropriate proportion of resources diverted to pupils on the borderline of achieving the target standard, to the detriment of both higher achievers and of those with little or no hope of reaching the target, even with assistance.¹¹⁴ Brian Lightman, President of the ASCL, told us that:

[...] if your target focuses on five grades A* to C, inevitably, the focus will be on those with four and who are nearly heading towards the fifth. You will concentrate on giving those children the extra help. [...] The [other] children [...] who do not quite fit into those categories, will be left out. That has been one of the major shortcomings of this target-setting culture over many years. For example, the focus of GCSEs has been very heavily on the C-D border line, and not, for example, on students underachieving by getting a grade A, but who could hopefully get an A*, or on those getting a B, but who could be helped to get an A.¹¹⁵

Even the Minister, Jim Knight, admitted that:

At the end of Key Stage 2 there is too much focus in some schools on people on the margins of a level 4 and not on the rest, because that is where the measure is.¹¹⁶

111 Q339

112 Smithers, A (2007) "Schools." In *Blair's Britain 1997-2007*. Ed Anthony Seldon, pp 361-384. Cambridge: University Press, 2007, 978-0-521-70946-0

113 Ev 122

114 Ev 265

115 Q160

116 Q374

The effect of concentrating on borderline pupils can be pernicious for the individual. The Association of Colleges stated that, whilst a pupil may have the necessary grades to progress to the next level, if that learning is shallow, focussed only on passing the test, they may not have a full grasp of the necessary concepts or sufficient intellectual rigour to deal with the demands of the next level. They conclude that “This raising of false expectations resulting in a sense of inadequacy may well account for the high drop out rate at 17”.¹¹⁷

75. Targets have had the effect of refocusing effort on maximising a school’s achievement, for which it is held accountable, rather than crediting the achievements of individual pupils. This is because, using the mechanisms of teaching to the test and narrowing the taught curriculum, it is possible to inflate test scores without improving the underlying education of the children taking those tests.¹¹⁸ The ATL cites a study which finds that a focus on the achievement of a particular level, together with booster classes, may have the effect of assisting pupils to achieve a Level 4 in mathematics, for example, but that this level is not sustained over a period of six months to a year.¹¹⁹ Professor Peter Tymms of Durham University told us that externally-imposed targets aimed at complex processes, such as teaching and running a school, are less effective than targets aimed at simple tasks which are easily quantified. In his view, targets for schools should come from within rather than being imposed externally.¹²⁰

76. Section 2 of *Making Good Progress* gives an indication of the Government’s view of attainment against targets. The national results for pupils in English at Key Stages 2 and 3 are set out, with Level 4 being the expected standard at Key Stage 2 and Level 5 the expected standard at Key Stage 3. In both cases, the 2008 target is that 85% of all pupils should reach the expected standard by the end of those Key Stages. The figures presented suggest that there is some distance to cover to reach these targets: in 2006, 79% of pupils achieved Level 4 or above at Key Stage 2; and in 2005, 74% of pupils achieved Level 5 or above at Key Stage 3. The Government states that those not meeting expected standards are “moving too slowly”, and that “It is disappointing that they have made such slow progress”. The language used to refer to these struggling pupils is that of “success” and “failure”, for example, “[...] by no means all of [those children] with SEN had severe neurological problems effectively preventing success”; and “[...] apart from severe neurological disorder, none of these characteristics necessarily results in failure”. But most parents want their children to make the best progress they are capable of making. We warn later (paragraphs 212-218) against developing a crude national progress standard which demands that every child progresses at the same rate, but nevertheless feel that a more subtle measurement of progress could act as the spur to improved teaching and learning.

77. When the assessment levels, for example Level 4 for 11 year-olds, were developed in the 1980s, they were the levels achieved by the average child. In the following years they have become a minimum expected standard, thus creating an expectation that improved teaching and learning can enable most pupils to achieve what used to be average levels. In

117 Ev 200

118 Smithers, A (2007) “Tests, tables and targets.” In *Tables, Targets and Testing—It’s Time to Trust Our Schools*. NAHT Commission of Inquiry into Assessment and League Tables, pp 50-61, Haywards Heath: NAHT.

119 Ev 60-61

120 Q27

this light, the Government's proposals in *Making Good Progress* that pupils should progress by two levels in one Key Stage is an even more challenging target and we are concerned that it runs counter to the Government's policy on personalisation.

78. The Government's point is, quite reasonably, that by no means all of those children falling short of expected standards are necessarily incapable of achieving them and that different teaching and learning strategies may improve their performance and help them to fulfil their potential. However, we think that the language of 'success' and 'failure' highlights a problem with the standards agenda which the Government's reasoning does not address. The NAHT challenged the Government's approach, pointing out that children learn at different rates and in different ways. Some children will easily surpass the expected standards at the end of a Key Stage and others will need much longer to reach them. Schools should focus on assisting children to reach the goals which are appropriate for them as individuals. The NAHT concludes by stating that:

We must not label as failures 11 year olds who learn more slowly or who have skills in different aspects which cannot be described in such concepts as "level 4". What is a level 4 Happiness or a level 5 Social Responsibility? How can we expect a certain, arbitrary percentage to succeed or fail? More importantly, why should we?¹²¹

79. Mick Brookes, General Secretary of the NAHT, stated that target-setting is of extreme importance to individual children and should not be controlled centrally by government:

If you set targets too high, the child cannot do it, becomes frustrated and disconnects. If you set that target too low, the child becomes bored and disconnects; they then leave school as soon as they can—24% of them. So target-setting is a very individual and personalised event.¹²²

The ATL similarly expressed concern that the perceived importance of targets, especially Level 4 at Key Stage 2, is so strong that many pupils who do not reach that level feel like failures. Other witnesses have made related arguments in relation to children with special educational needs. It has been argued that their 'failure' to meet national targets leads to them being marginalised and devalued.¹²³ Worse, because of the way school accountability is tied to test results for each school, these individuals are seen as a burden on the school itself as they drag down its aggregated test scores.¹²⁴

80. The NUT drew our attention to the EPPI (Evidence for Policy and Practice Information) study (2004), on the impact of repeated testing on pupils' motivation and learning. The review concluded that repeated testing and examination demotivated pupils and reduced their learning potential, as well as having a detrimental effect on educational outcomes. Other key findings included evidence which showed that teachers adapt their teaching style to train pupils to pass tests, even when pupils do not have an understanding

121 Ev 71

122 Q161

123 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1(c)

124 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 2(b)

of higher order thinking skills that tests are intended to measure and that National Curriculum tests lower the self-esteem of unconfident and low-achieving pupils.¹²⁵

81. We endorse the Government's view that much can and should be done to assist children who struggle to meet expected standards. However, we are concerned that the Government's target-based system may actually be contributing to the problems of some children.

82. We believe that the system is now out of balance in the sense that the drive to meet government-set targets has too often become the goal rather than the means to the end of providing the best possible education for all children. This is demonstrated in phenomena such as teaching to the test, narrowing the curriculum and focussing disproportionate resources on borderline pupils. We urge the Government to reconsider its approach in order to create incentives to schools to teach the whole curriculum and acknowledge children's achievements in the full range of the curriculum. The priority should be a system which gives teachers, parents and children accurate information about children's progress.

Performance tables

83. The national test results for each school are aggregated into performance tables, published by the DCSF, which allow comparison of school results, either in terms of absolute test scores or in terms of Contextual Value Added scores (see further paragraphs 90-102). The tables also allow for instant comparison of a school's results against both the local authority average and the national average. A limited amount of background information is given for each school, including the number of pupils on the school roll and the number and percentage of pupils with special educational needs ("SEN").¹²⁶ The DCSF does not rank schools in order of achievement although, as we have mentioned above (paragraph 63), ranked tables are compiled and published by the news media. Although some have argued that performance data should not be published in tables, others have countered that, if the Government did not do it, the media would.¹²⁷ The Minister told us that he thought it was better that this data should be published in a controlled and transparent manner by the Government, rather than leaving publication to the media. He thought that there would be "an outcry that we were hiding things" if the Government did not publish performance data on schools.¹²⁸

84. One of the major criticisms of performance tables is that they do not provide a true reflection of the work performed in schools.¹²⁹ A significant aspect of this argument relates back to the discussion in Chapter 2 on the purposes of testing. If national tests are primarily designed to measure pupil attainment at particular points in their school careers¹³⁰, can they be said to measure the performance of schools with a sufficient degree

125 Ev 263

126 <http://www.dcsf.gov.uk/performancetables/>; Ev 158

127 Ev 240-241; Q13; Q14

128 Q336

129 Ev 69; Ev 70; Ev 83; Ev 115; Ev 198; Ev 227; Ev 244; Ev 247-248; Q153; written evidence from Lorraine Smith, Headteacher, Western Church of England Primary School, Winchester, para 13

130 Q79

of validity? The answer to this question depends on what is being measured. If school performance is viewed purely and simply in terms of its pupils getting the highest possible marks, especially in comparison with other, similar schools, then test results might be a valid measure of school performance. If, on the other hand, school performance is seen as rather broader than this, to include teaching of a full and rounded curriculum, artistic, cultural and sporting activities and good pastoral care, as well as academic achievement, then aggregated test results are not a valid measure of school performance, because they only measure one, narrow aspect of it.¹³¹ As the Advisory Committee on Mathematics Education stated:

Clearly the league tables only measure what is tested and not wider aspects of school performance.¹³²

85. The NAHT similarly argued that:

Even modified by social deprivation or value added factors, [performance tables] can only give a distorted snapshot of the work of a vibrant and organic community.¹³³

86. We were given an extreme example of this by Mick Brookes, General Secretary of the NAHT. He told us about a school which persistently languishes at the bottom end of the performance tables but whose Ofsted report paints a picture of excellent leadership, good improvement in all areas of school life and a strong commitment to the personal development of all pupils. Mr Brookes continued:

There are very good schools that work against the odds to produce higher educational qualifications than they have ever had in their areas, but they are disabused of that excellent work on an annual basis.¹³⁴

87. The NAHT stated that, although national tests provide statistically valid results at a national level, individual school cohorts may be small in size, with a single pupil counting for more than 15% of the overall score for a school. Statistically, this means that the ranking of a small school according to data in the performance tables can be extremely volatile from year to year, depending on the cohort being tested.¹³⁵ The NAHT therefore calls for care in the interpretation of test results at school level and argues that test data should be used as one indicator, alongside many others, of school performance, rather than as “a determinant”.¹³⁶

88. The GTC argued that parents’ desire for information about their children’s learning and progress will not be best served by the single measure of Key Stage test results. The information presented in the performance tables do not tell parents and the local community the full story about the education taking place in a school. The GTC cited

131 Written evidence from Advisory Committee on Mathematics Education, para 25; written evidence from Lorraine Smith, Headteacher, Western Church of England Primary School, Winchester, para 13

132 Written evidence from Advisory Committee on Mathematics Education, para 25

133 Ev 69

134 Q138

135 Q13

136 Ev 68; see also written evidence from Association for Achievement and Improvement through Assessment, para 7

evidence which indicates that, when parents make judgements about the quality of a school, they do not use the school's position on published league tables as their main criterion (GfK NOP Social Research 2005). Parents, it argues, require broadened and enriched sources of information about their local schools.¹³⁷ Keith Bartley, the Chief Executive of the GTC, told us that MORI research in 2005 showed that parents attributed low value to performance table data when determining their choice of school partly because they found the information confusing.¹³⁸ The Minister disagreed with this view, stating that "They are very simple and easy to understand".¹³⁹

89. We endorse the view, put forward by many witnesses, that the data presented in the performance tables give only a very limited picture of the work which goes on in a school. It is, therefore, never appropriate to rely on this information alone when forming a judgment about a school's overall performance.

90. Comparison of schools based on raw test scores is, we think, self-explanatory. We shall, however, explore in a little more detail the use of Contextual Value Added scores ("CVA") and what they mean. CVA scores at school level are, essentially, a measure of progress over time from a given starting point and they provide a means of assessing the relative effectiveness of a school. Unlike the raw test scores presented in the performance tables, it is not a measure of absolute attainment. The period of the progression measurement will be given for each score, for example Key Stage 1-2 or Key Stage 2-4. Although CVA is a measure of progress based on prior attainment, it is adjusted to take account of a variety of factors which have been statistically shown to have an impact on an individual's progress.

91. The first step is to calculate a CVA score for each pupil. These scores are then used to calculate the CVA score for the school as a whole:

The pupil calculation

1. Calculate a prediction based on the pupil's prior attainment (eg based on an average of Key Stage 2 results for English, maths and science).
2. Adjust this prediction to account for the particular characteristics of the pupil (eg gender, ethnicity, eligibility for free school meals, degree of deprivation).
3. Adjust further to take account of prior school-level attainment.
4. Calculate the CVA score for the individual pupil by measuring the difference (positive or negative) between the pupil's actual attainment and that predicted by the CVA model.

The school calculation

Simply put, a **pupil's** CVA score is effectively the difference between a set of results predicted according to the CVA model and the results actually achieved. A school adding

137 Ev 75

138 Q154

139 Q338

value will enable the pupil to outperform the results expected of him or her according to the CVA model. In order to calculate the CVA score for a **school** as a whole, the average CVA scores of its pupils is taken, then an adjustment is made for the number of pupils in a school's cohort (the shrinkage factor).

92. This provides a measure of school effectiveness. The Government said that:

CVA is a powerful tool for analysing school performance, but it is a relative measure, only effective when seen in combination with other factors, including raw scores, value added based on prior attainment, school self-evaluation, inspectors' judgements and the content of the school profile. It shows a school's past performance in relation to other similar schools, but cannot project performance into the future. Hence it cannot be used to set future targets.¹⁴⁰

93. The CVA measure is only an estimate, even within its own, limited terms. The score is based on a prediction which is, in turn, based on the actual attainment of a pupil in a given exam on a given day. On another day with the same pupils, a school may well have achieved somewhat different results. This degree of uncertainty is reflected in the confidence interval, which is provided in the performance tables alongside the school's CVA score. The confidence interval is, essentially, the range of scores within which one can be statistically confident that the "true" school effectiveness (according to the model, at least) will lie. It gives a measure of the uncertainty inherent in a school's CVA score and the size of the confidence interval will be determined by the number of pupils in the calculation.

CVA information for each school is presented on the DCSF website, illustrated by the following example:

	Contextual Value Added				
	KS1-2 score	CVA	Coverage	CVA Confidence Intervals	
				Lower limit	Upper limit
LA Average	100.1		95%		
School A	99.5		97%	98.8	100.2
School B	99.9		98%	99.5	100.4

Source: DCSF website; Achievement and Attainment Tables.

94. The major issue which arises out of CVA is, as with performance tables generally, that it is not readily understandable to the layman, and parents in particular. It is not clear from the table above what the practical difference is between a school with a Key Stage 1-2 score of 99.5 and a school with a score of 99.9. In fact, other tables available on the DCSF website show that the absolute results for School A are low, whereas the absolute results for School B are high. The implication of the CVA scores is that both schools are similarly effective, albeit with very different intakes (School A has a high number of SEN pupils, School B a

relatively low number, according to yet another table). However, this interpretation is not obvious unless one undertakes a thorough analysis and comparison of several tables together. It follows that the intended use of CVA scores, to place the absolute results in context, is diluted because many do not know how to interpret them.

95. We put this concern to the Minister. He told us that:

I do not think that it is that difficult to understand that in CVA terms, 1,000 is the norm. If you are above 1,000, you are adding value better than the norm. If you are below 1,000, you are adding value lower than the norm. If that is all people understand, then it is pretty straightforward.¹⁴¹

Mr Tabberer added that, in publishing performance tables, including CVA scores, the Department is “following the principle of being transparent about all of the analyses so that parents can access the information that they understand or the information that they want”. He said that the Department tries to ensure that the public can see comparators and benchmarks and that CVA scores were considered a fair means of comparison.¹⁴² We do not take issue with this, although we have already noted the limitations of any kind of evidence based on test scores alone. We cannot agree, however, that the meaning of CVA scores, as they are presented in the Department’s own performance tables, is by any means obvious.

96. We consider that CVA scores are important because there is a strong correlation between the characteristics of a school’s intake population and its aggregated test results and CVA attempts to make some compensation for this. Schools with an intake of lower-performing pupils will do less well in the performance tables of raw scores than schools with an intake of higher-performing pupils.¹⁴³ Mick Brookes put it bluntly, claiming that performance tables simply indicate “where rich people live and, sadly, where poor people live as well”.¹⁴⁴ Dr Mary Bousted, General Secretary of the ATL, told us:

Variation between schools is much less than variation within schools. We know a school can make about 14% of the difference; the rest of the determining factors on a child’s achievement come from their background, actually. And 14% is a lot; I am not undermining what a school can do. However, that means that schools in the most challenging areas have to work extremely hard to get the results that they do.¹⁴⁵

97. CVA scores, then, go some way towards levelling out these inequalities. However, they are not a transparent measure, and it is not easy to judge the validity of the variables used.¹⁴⁶ CVA is still a relatively blunt instrument for making comparisons, based as it is on the limited dataset of test results and on a series of assumptions and generalisations about a pupil’s background. Nigel Utton, Chair of Heading for Inclusion, illustrated this point:

141 Q360

142 Q357

143 Ev 52; Ev 115; Ev 201; Q22; Q23; written evidence from LexiaUK, para 2.7; written evidence from Advisory Committee on Mathematics Education, para 25

144 Q146

145 Q154

146 Ev 115

In the case of my own school, by removing two children with statements for educational needs from the statistics we move from being significantly below the national average to being within normal boundaries. If the contextualised value added measures were sufficiently sophisticated, that would not be possible, as it would weight children with SEN to factor out such a discrepancy.¹⁴⁷

98. Whilst we consider that Contextualised Value Added scores are potentially a valuable addition to the range of information available to parents and the public at large when making judgments about particular schools, we recommend that the information be presented in a more accessible form, for example graphically, so that it can more easily be interpreted.

99. We are concerned about the underlying assumptions on which Contextualised Value Added scores are based. Whilst it may be true that the sub-groups adjusted for in the Contextualised Value Added measure may statistically perform less well than other sub-groups, we do not consider that it should be accepted that they will always perform less well than others.

100. In addition to these specific recommendations about Contextual Value Added scores, we recommend that the Government rethinks the way it publishes the information presented in the Achievement and Attainment Tables generally. We believe that this information should be presented in a more accessible manner so that parents and others can make a holistic evaluation of a school more easily. In addition, there should be a statement with the Achievement and Attainment Tables that they should not be read in isolation, but in conjunction with the relevant Ofsted report in order to get a more rounded view of a school's performance and a link to the Ofsted site should be provided.

101. We have received some evidence that Ofsted places considerable weight on test scores when making judgments about schools under the new, lighter touch, inspection regime. The NUT said that Ofsted relies on CVA as the baseline measure for school evaluation.¹⁴⁸ Heading for Inclusion similarly stated that Ofsted inspections:

[...] focus almost entirely on a school's ability to produce high results in tests at various stages, whether they be Key Stage SAT results or GCSEs. This has led to schools devoting much of their time to 'playing the game' and teaching the children to pass the tests.¹⁴⁹

Other witnesses also expressed concern that Ofsted uses information in the performance tables as key inspection evidence.¹⁵⁰ Cambridge Assessment made the point that the new Ofsted inspection regime is far more dependent on national assessment data than previously. Although Cambridge Assessment stated that the new regime has been broadly

147 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 2(b)

148 Ev 262

149 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1(d)

150 Ev 48; Q153; Q169; written evidence from Advisory Committee on Mathematics Education, para 3; written evidence from Lorraine Smith, Headteacher, Western Church of England Primary School, Winchester, para 7

welcomed by schools, it argued that the regime fails to take into account the essential weaknesses in these data.¹⁵¹ The IPPR gives a measured assessment of the situation:

[...] the results of national tests are a critical input into Ofsted inspections, and a bad inspection may result in a school being issued a notice to improve, or risk being placed in special measures. Entering special measures means that a school loses its autonomy and represents a severe criticism of the leadership of the school. [...]

It is quite right that there should be a robust inspection mechanism to provide schools with powerful incentives to improve, and especially to ensure that no school falls below a minimum acceptable standard. However, if test results are to play an important role in such a powerful incentive mechanism, it is all the more important that they are robust, valid, and do not negatively impact on other desirable aspects of the learning environment.¹⁵²

The IPPR added, however, that it is important not to overstate these arguments and that Ofsted does take into account a wide range of other factors in its inspections.¹⁵³

102. The scope of this inquiry does not extend to a thorough examination of the way Ofsted uses data from the performance tables under the new, lighter touch, inspection regime. However, we would be concerned if Ofsted were, in fact, using test result data as primary inspection evidence in a disproportionate manner because of our view that national test data are evidence only of a very limited amount of the important and wide-ranging work that schools do.

103. So far, we have considered objections to performance tables, including CVA measures, based on arguments that they are not a valid measure of the performance of schools judged across the full range of their activities; and they are not readily understandable by those who may wish to use them, especially parents. However, the most serious and widespread objection to performance tables is, as with performance targets, the distorting effect that they have on the education which takes place in schools.

104. The Government states that the performance tables are “an important source of public accountability for schools and colleges”.¹⁵⁴ The use of performance tables for school accountability means that a school’s standing in the performance tables is a matter of significant importance to that school, directly or indirectly affecting the morale of pupils and teachers; the attitudes of parents; the school’s standing in the local community and within the wider local authority; the resources allocated to it; and perhaps even the school’s very survival. The stakes, as many witnesses have pointed out, are high.¹⁵⁵

105. The evidence we have received overwhelmingly suggests that these high-stakes lead to serious distortion of the education experience of pupils (and see further Chapter 4): teaching to the test, narrowing of the taught curriculum and disproportionate focus on

151 Ev 216

152 Ev 238

153 Ev 238

154 Ev 158

155 Ev 58; Ev 65; Ev 70; Ev 200; Q224

borderline students.¹⁵⁶ Witnesses have commented that the use of performance tables as accountability measures has had the effect of “undermining good practice in many classrooms”¹⁵⁷ and has encouraged a “risk-averse culture”¹⁵⁸. Performance tables “depress and demotivate teachers who struggle to make children achieve grades they are not quite ready for”.¹⁵⁹ The NASUWT told us that the practical effect of performance tables:

[...] is to contribute to a skewing of the curriculum, generate unacceptable levels of pressure and workload at school level and entrench a competitive rather than collaborative culture between schools. They are also responsible for many of the pressures that inhibit the ability of teachers to exercise an appropriate level of professional discretion and autonomy.¹⁶⁰

Professor Tymms argued that:

We are forcing teachers to be unprofessional. League tables are an enemy of improvement in our educational system, but good data is not.¹⁶¹

106. We consider that schools are being held accountable for only a very narrow part of their essential activities and we recommend that the Government reforms the performance tables to include a wider range of measures, including those from the recent Ofsted report.

107. We have considered in this Chapter some of the issues which arise from the use of national test results for the purposes of accountability and monitoring of schools through performance targets and tables. In the next Chapter, we shall consider in more detail some of the unintended consequences of this regime and suggestions for radical reform of the accountability system.

156 Ev 57; Ev 58; Ev 65; Ev 70; Ev 83; Ev 246; Ev 247; Q224; written evidence from Doug French, University of Hull, para 2.4; written evidence from Advisory Committee on Mathematics Education, para 24; written evidence from Mathematics in Engineering and Industry; written evidence from Institute of Physics, para 6; written evidence from Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hants; written evidence from Association of Science Education, para 18

157 Written evidence from Mathematics in Engineering and Industry

158 Ev 57

159 Written evidence from Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hants

160 Ev 246

161 Q13

4 The consequences of high-stakes uses of testing

108. In previous chapters, we have alluded to the concept of high-stakes uses of testing and a variety of unintended consequences, including distortions of the education experience of pupils, which are said to result. In this Chapter, we shall consider, first, what is meant by “high-stakes” and then examine in more detail the claims which are made about the consequences resulting from high-stakes testing.

109. The NAHT argues that the tests themselves are not inherently problematic, but the use of the results of those tests for high-stakes purposes is. NAHT members do not take issue with the principle of testing, but with the emphasis on published performance tables and the links between test results and inspection outcomes.¹⁶² The QCA agrees that, when evaluating an assessment system, there is a need to distinguish “the impacts attributable to testing, *per se*, and the impacts attributable to the high-stakes uses of results which the testing is designed to support”.¹⁶³

110. Ralph Tabberer told us that he questioned the premise that Key Stage tests were high-stakes in conventional terms. In his view, high-stakes tests were those, such as the 11-plus, which determine which school pupils will attend within a selective system. He considers that Key Stage tests are “medium-stakes”, allowing pupils to demonstrate their attainment and giving them and their parents a sense of their level of achievement. It is incidental, according to Mr Tabberer, that Key Stage tests “also happen to give us very useful information [...] for policy development and accountability”.¹⁶⁴

111. We think that the stakes of the national testing system are particularly high for schools and teachers at all levels and for young people at 16+ who need qualifications. Children need to do well in tests and, later on, 16+ qualifications, in order to move successfully on to the next level of schooling and to get the grades they need for their chosen higher education course or employment. Teachers and headteachers need their pupils to do well in tests in order to demonstrate that they are effective teachers, to win promotion and, perhaps, financial reward. Perceived failure of teachers in this respect may lead to demoralisation, being passed over for promotion, difficulty in finding employment in other schools or leaving the profession altogether.¹⁶⁵ Schools are held accountable according to the test results achieved by their pupils and a poor set of results may result in public humiliation in published performance tables, being perceived as a “failing school”, interventions by Ofsted and even closure in extreme cases.¹⁶⁶ Local authorities may also be perceived as “failing” if the schools in their area do not demonstrate results measuring up to the expected standards. Finally, the Government itself, having put in place accountability mechanisms for driving up standards from the centre, stands to lose a considerable

162 Ev 70

163 Ev 29

164 Q346

165 Ev 55; written evidence from Advisory Committee on Mathematics Education, para 3

166 Ev 234

amount of political capital if targets are not met and standards are not seen to be improving according to the measures set out in performance tables.¹⁶⁷

112. When “high-stakes” are mentioned in the evidence we have received in the course of this inquiry, the term most often refers to high-stakes for the school and teachers, rather than children. This is because it is the use of test results for the purpose of school accountability which is blamed for rendering the tests high-stakes in the first place. As should be apparent from the previous Chapter, it is the link between test results, performance targets and performance tables which raises the stakes for schools and, therefore, teachers. Under the current testing regime, pupil performance in tests is inextricably linked with school accountability across the board. It seems relatively clear to us that, even if this link is broken, national tests and public examinations will remain high-stakes for individual children because their own futures are at stake. In fairness to schools and teachers, their objection to the current, high-stakes accountability regime appears to us to be as much because it distorts the education experience of children as out of self-interest. Jerry Jarvis, Managing Director of Edexcel, told us that:

Unfortunately, there are huge issues at stake in most schools, and teachers are human. Having said that, the huge overwhelming majority of teachers aim to deliver on education—that aim comes across strongly in what they do. However, there is no question that there is pressure.¹⁶⁸

113. However, we have received considerable evidence of problematic practice, illustrated by this account from a teacher:

Last year my Headteacher asked me for my reaction to the KS3 Maths results so I started talking about the progress of individual students and how pleased I was for them. He brushed these comments aside, merely wanting to talk about percentages at each level and comparisons with other year groups/other schools.¹⁶⁹

Clearly, there are serious problems which need to be addressed. The following sections will analyse some of these issues in more detail.

Teaching to the test

114. AQA, an Awarding Body accredited by the QCA, identifies that:

There is a clear tension between the need to ensure that schools are properly accountable on the one hand and the need to allow room for curriculum innovation and inspired teaching on the other.¹⁷⁰

It is this tension which is apparent in any discussion of “teaching to the test”. Essentially, teaching to the test amounts to teachers drilling their pupils in a subject on which they will face a test or examination. In extreme cases, a high proportion of teaching time will be

167 Ev 60

168 Q193

169 Ev 46

170 Ev 102

given over to test preparation. The focus of lessons will be narrow, with teachers coaching their pupils on examination technique, question spotting, going over sample questions similar to those likely to be set in the test and generally focussing teaching of the substance of a subject in a way best calculated to maximise marks in the test. The IPPR identifies a range of classroom practices which are intuitively of low educational value but which may, nevertheless, improve test results, including narrow learning, where teachers concentrate on aspects of the curriculum likely to be tested; shallow learning, where teachers focus on the way in which a component of the curriculum is likely to be tested; question spotting; and risk-averse teaching with low levels of innovation.¹⁷¹

115. The DfES stated in its memorandum its view of what should be happening in classrooms. Children should acquire deep knowledge and understanding of a concept through extended experience and practice. At the same time, they should certainly have an appreciation of what examiners are looking for and how to present their answers in a test. Preparation for tests should be wholly integrated into the classroom experience, with teachers and pupils agreeing targets for the next stage of learning and discussing what the pupil needs to do to reach those targets. Progress should be tracked continuously through assessment, benchmarked periodically with formal tests such as the optional tests provided by the QCA. There should be no cramming in Years 6 and 9 as the Key Stage tests approach.¹⁷² The DfES concluded that:

The teacher who prepares pupils for a test without developing the deeper understanding or more extended experience required may be fortunate enough to enjoy some short-term success, but will not be likely to maintain that performance over time.¹⁷³

However, as Key Stage 2 (and often Key Stage 1) testing occurs as a child leaves one institution to go to another, this in practice is not a problem for the teacher who is responsible for teaching the child who takes that test. This increases the need for the government to find alternative measures of the effectiveness of teaching which encourage teachers to ensure that learning achieves more than a short term ability to pass a test.

116. The Government's statement of what *should* be happening (paragraph 115) seems to us rather out of touch with what appears to be happening in classrooms according to the evidence we have received. It has been argued by a great number of witnesses to this inquiry that the high-stakes attached to national testing increases pressure to teach to the test. The IPPR noted that it is difficult to prove this causal link¹⁷⁴; and the QCA noted that there is little, systematically documented evidence of the phenomenon, only considerable anecdotal evidence¹⁷⁵. However, the vast majority of other witnesses have had no such reticence: they are clear that teaching to the test happens, that it is prevalent and that it is caused by the high-stakes (for schools) which are attached to the results of the tests.¹⁷⁶

171 Ev 239

172 Ev 160

173 Ev 160

174 Ev 239

175 Ev 30

176 Ev 48; Ev 57; Ev 59; Ev 60; Ev 68; Ev 69; Ev 73; Ev 83; Ev 115; Ev 199; Ev 200; Ev 217; Ev 224; Ev 227; Ev 269; Ev 271; Q128; Q139; Q169; Q268; written evidence from Heading for Inclusion, Alliance for Inclusive Education, paras 1(c) &

Innovation and creativity in teaching approach is considered “too risky”¹⁷⁷ and teaching to the test is displacing healthy classroom practice, such as informal, formative teacher assessment.¹⁷⁸ The ASCL’s view is typical of the evidence we have received:

Teachers have been criticised for teaching to the test but, if the system is geared to constantly monitoring progress and judging teachers and institutions by outcomes, it is hardly surprising that the focus is on ensuring that students produce the best results.¹⁷⁹

117. The effect of high-stakes on teachers appears to be profound. Mick Brookes gave the example of a hypothetical young headteacher or deputy head with a young family and a mortgage:

[...] a system of fear has been inculcated throughout education [...] You do not want to go to headship because you know that you will carry the can. Unfair and unfounded decisions are made on the performance of schools because Ofsted is now relying far too heavily on the data that we have discredited during this presentation. It is having a profound effect not only on the curriculum, but on the recruitment and retention of head teachers, in particular, who carry this can and do not survive being put into a category, on many occasions quite unfairly.

118. Cambridge Assessment identified teaching to the test as “a very serious issue” which may be one significant factor, although not the only one, in the so-called ‘plateau-effect’ which has been associated with “the majority of innovations such as the Primary Literacy and Numeracy Strategies”. As Cambridge Assessment puts it, “A succession of well-intended and seemingly robust initiatives repeatedly run out of steam”.¹⁸⁰ The phenomenon is especially prevalent as the time for the test approaches. The NAHT refers to a recent survey which indicates that, in Year 6, for four months of the school year teachers are spending nearly half their teaching time preparing pupils for Key Stage 2 tests.¹⁸¹ It has been argued that in Wales, where testing is no longer high-stakes, healthy practice in classrooms, such as cross-phase moderation, is now being adopted.¹⁸²

(d); written evidence from Portway Infant School; written evidence from Doug French, University of Hull, para 2.1; written evidence from LexiaUK, para 2.8; written evidence from Advisory Committee on Mathematics Education, paras 1 & 26; written evidence from Campaign for Science and Education, para 24; written evidence from Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hants; written evidence from The Royal Society, para 5; written evidence from Association of Science Education, paras 24-25; written evidence from The Mathematical Association; written evidence from S Forrest, Teacher of Mathematics, Wokingham, para 9; written evidence from Science Community Partnership Supporting Education, section 2; written evidence from Warwick Mansell

177 Ev 113; Ev 224; written evidence from Association of Science Education, para 22; written evidence from Doug French, University of Hull, para 1.3

178 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 2(c)

179 Ev 48

180 Ev 217

181 Ev 69

182 Written evidence from The Mathematical Association

Shallow learning

119. One serious consequence of teaching to the test is that it tends to lead to shallow learning and short-term retention of knowledge.¹⁸³ The Mathematical Association pointed to “a narrow focus on mark-winning behaviours rather than teaching [pupils] a coherent understanding of the subject”.¹⁸⁴ Another witness had said that the positive changes to the science curriculum have been undermined by “a system which values factual recall and superficial conceptualisation over deeper understanding and engagement”.¹⁸⁵ Hampshire County Council states that it has identified widespread teaching to the test, leading to a reduction in the time spent exploring more imaginative and creative aspects of the curriculum and an emphasis on short-term memorisation and ‘test tactics’ rather than deep learning and understanding.¹⁸⁶ In an evaluation study of 14–19 mathematics teaching, Ofsted identified:

A narrow focus on meeting examination requirements by ‘teaching to the test’, so that, although students are able to pass the examinations, they are not able to apply their knowledge independently to new contexts, and they are not well prepared for further study.¹⁸⁷

120. The ATL also referred to research which found that teaching to the test led to many pupils not possessing the skills or understanding which the test was designed to assess and that “the focus of teaching in this environment is to teach students to pass tests even where they do not have the skills or understanding”.¹⁸⁸ The reason is that tests do not usually test the full range of what is taught. When the tests are high-stakes for schools, the pressure on schools and teachers is to focus on pupils’ performance in those areas most likely to be tested. In the context of an overburdened curriculum, those areas will dominate classroom time. Where resources are focussed in this way, the breadth and depth of subject-coverage suffers.¹⁸⁹ The ATL considers that “the purpose of assessment as an aid to the development of learning is shunted into second place”, with maximising test results to enhance accountability measures promoted to first place.¹⁹⁰

121. This is extremely worrying and may provide a partial explanation for the apparent decline in attainment of pupils moving from Year 6 into Year 7.¹⁹¹ The NAHT states that:

There are many examples of year 6 students who have obtained high levels, particularly in Science SATs, who are not able to replicate this performance within

183 Ev 226; Q128; written evidence from Association of Science Education, para 22; written evidence from Institute of Physics, para 3; written evidence from Campaign for Science and Education, para 23; written evidence from Association of Science Education, para 20

184 Written evidence from The Mathematical Association

185 Written evidence from Science Community Partnership Supporting Education, para 1

186 Ev 271

187 Ofsted, *Evaluating Mathematics Provision for 14-19 Year Olds*, May 2006, <http://www.ofsted.gov.uk/assets/4207.pdf>

188 Ev 60

189 Ev 60

190 Ev 60

191 Q139; Written evidence from Advisory Committee on Mathematics Education, para 27; written evidence from Association for Achievement and Improvement through Assessment, para 11; written evidence from S Forrest, Teacher of Mathematics, Wokingham, para 9

the secondary curriculum. The results are not wrong. They merely indicate that the students have learned how to pass Science SATs and not developed scientific skills and absorbed scientific content. This can be extremely unhelpful for the receiving secondary school.¹⁹²

122. The QCA reports that schools often mistrust the results from the previous Key Stage and re-test using different measures.¹⁹³ The Chartered Institute of Educational Assessors suggests that the Year 6 to Year 7 dip may, indeed, be caused by pupils being hotheaded and studying a limited curriculum in the final term of primary education in preparation for the Key Stage 2 tests. On arriving in a new institution, there is no immediate public examination in prospect, so pupils are inclined not to work as efficiently or effectively. There, however, are other factors at play. Pupils are also moving from a regime in which they were taught by the same teacher for all subjects, to a regime in which “they are taught by specialist teachers using specialist equipment in discrete physical locations for each curriculum area”. They are with new peers in a new social environment in which they are now the youngest, having been the oldest and top of the hierarchy in their primary schools.¹⁹⁴ It is also worth noting that, in Science, the Key Stage 3 curriculum is much broader than the Key Stage 2 curriculum so that, although a Level 4 at Key Stage 2 might be equivalent in difficulty to a Level 4 at Key Stage 3, it is not equivalent in scope.

Qualifications

123. In qualifications, the Awarding Bodies arguably encourage teaching to the test through factors such as increasingly more closed questions, the provision of sample questions and answers and of extensive syllabus training to teachers, including comprehensive teaching and learning materials.¹⁹⁵ The AQA prides itself on its extensive programme of “teacher support”, including support at the beginning of a new syllabus, regular review opportunities, rapid feedback from examiner reports and the provision of “comprehensive, innovative and motivating teaching and learning materials” to schools selecting AQA as their examination provider. The provision of teaching and learning support is, apparently, a major and growing part of AQA’s work.¹⁹⁶ The AQA’s Deputy Director General, Dr Andrew Bird, thought that the provision of curriculum support materials and teacher training was important, but emphasised that exams were set by a separate part of the organisation and that it guards against predictability.¹⁹⁷ Greg Watson, Chief Executive of OCR, told us that:

By being a bit more open, a bit more transparent, and providing a few more clues, we enable people to feel well prepared, and what we are actually [doing] is assessing what they can do not how successfully they have guessed what they are about to do, or their ability to cope with the surprise of what they have been faced with. I think that that has also been a positive development. But I would set against that the fact

192 Ev 69

193 Ev 32

194 Ev 227

195 Ev 102; Q228

196 Ev 102

197 Q181

that there is a challenge and we employ expert people in the field of assessment to make sure that we keep the level of challenge right, and that it does not become too formulaic and too predictable, which obviously would mean beginning to lose some of that effect.¹⁹⁸

124. Others take a different view of these practices. Warwick Mansell refers in his memorandum to his observation of two senior examiners giving teachers “tricks on how to boost results” and sees this as evidence of “the cynical lengths to which test preparation and the search for shortcuts to improve pass rates can be taken”.¹⁹⁹

125. In addition to the training and materials detailed above, electronic tools from the Government (RAISEonline) and Edexcel (ResultsPlus) make available to schools data on test results which can be broken down by question, by pupil, by teacher, by year group and by school.²⁰⁰ City & Guilds favours this approach, stating that, “Considering the significant effort that goes into the final examining process by all parties the current under-use of this data is a travesty” and welcomes products such as Edexcel’s ResultsPlus.²⁰¹ AQA may be piloting a similar system this Summer, although OCR apparently has no plans along these lines.²⁰²

126. These data are intended to enable schools, teachers, parents and pupils to track the progress of individual pupils and year groups, to evaluate the effectiveness of teaching and to identify areas of strength and weakness.²⁰³ However, there is suspicion that the data will encourage further teaching to the test.²⁰⁴ Edexcel counters that “the genie [is] out of the bottle” and that trials of the system demonstrate that it can improve teaching and grades quickly and shows headteachers how well syllabuses are being taught.²⁰⁵ Jerry Jarvis told us that the system actually allows teachers to spend less time on revision and more time on true, personalised learning, tailored to the needs of an individual child.²⁰⁶ Greg Watson thought that this was nothing new and that teachers have been able, for some time now, to access examination answers and see how their pupils performed.²⁰⁷

Is teaching to the test detrimental?

127. It has been argued, especially by the Awarding Bodies, that teaching to the test is no bad thing if the tests assess a curriculum and are worth teaching to.²⁰⁸ The Government has largely avoided this issue and has not provided a definitive statement one way or the other, apart from its statement of what it deems to be proper classroom practice, set out in

198 Q181

199 Written evidence from Warwick Mansell

200 Ev 117-118; Ev 159

201 Ev 111; Q243

202 *Exam details get personal*, Times Educational Supplement, 10 August 2007

203 Ev 117-118; Ev 159

204 *Comparative data could benefit all*, Times Educational Supplement, 10 August 2007

205 *This genie of technology can help to boost attainment*, Times Educational Supplement, 10 August 2007

206 Q255

207 Q228

208 Ev 115; and, for example, Q224; Q292

paragraph 115 above. In his evidence to us, the Minister more or less rehearsed this statement, adding that “the vast swathe of teachers and schools ... use tests appropriately”. He added that £150 million was being invested over three years “In order to help those who do not [use tests appropriately] and to improve best practice generally”.²⁰⁹ The Minister did not have any statistics on the average amount of time spent in schools on preparing for tests, a situation which we find surprising considering the seriousness of the issues at stake and the strength of his assertions, and those of his officials, that teaching to the test is not a problem. When we asked him about the amount of time being spent by Key Stage 2 pupils on revision for the Key Stage 2 tests, the Minister replied that he saw nothing wrong with children learning what they needed to learn to pass the tests.²¹⁰ Ralph Tabberer also thought that revision was not “wasted time” and that it was important that pupils were prepared for Level 4 in order that they could access the secondary curriculum.²¹¹ Mr Tabberer did state, however, that:

[...] we do not want to see children being drilled so that they can just repeat low-level processes accurately and get marks for that—we are all clear that we do not want that.²¹²

He went on to say that, when he talks to teachers, he does not hear that they are drilling pupils, but that they are preparing them so that they can do their best in the tests. He thought that there was a “good balance”.²¹³ David Bell and Jon Coles took a similar view.²¹⁴ David Bell in particular thought that claims made about teaching to the test were overblown and did not match his experience, having visited hundreds of schools.²¹⁵

128. In reply, other witnesses have repeatedly pointed to the narrow range of knowledge and skills tested and the unreliable nature of test outcomes.²¹⁶ Warwick Mansell of the Times Education Supplement told us that teaching to the test cannot possibly be a positive phenomenon. The Awarding Bodies, for example, are becoming quite explicit about what is going to be in the examinations and publishes detailed marking schemes. He states that, as a result, “Pupils are being rewarded for dogmatic rule-following”, a situation which will not help to develop them as independent thinkers. Predictability of the examination content is, according to Warwick Mansell, the enemy of in-depth study of a subject. He believes that, although teaching to the test may always have been a feature in education, it is now far more prevalent due to the pressures on schools and teachers to raise results “more or less come what may”. He concludes that teaching to the test is fundamental to the learning experience of children and that “it is changing dramatically the character of education in this country”.²¹⁷

209 Q329

210 Q370

211 Q370

212 Q371

213 Q371

214 Q291; Q294; Q296

215 Q296

216 Q128

217 Written evidence from Warwick Mansell

129. The effects of this approach are felt long after children have finished their school education. Professor Madeleine Atkins, Vice-Chancellor of the University of Coventry, told us that teaching to the test and a test mentality on the part of students arriving at university leave them unprepared for the rigours of higher education. Students, particularly on vocational courses, arrive at university having learned techniques and how to apply them by rote. The consequent lack of deep understanding of the subjects they have studied at school leaves them unable to solve problems in real-world situations. Professor Atkins said that students find the transition to higher education difficult:

It does not mean to say that that they cannot do it, but it does mean that we have to teach in a rather different way to begin with in order that that synoptic understanding is developed and that understanding of connections between tools, techniques and methodologies is really in place.²¹⁸

Professor Steve Smith, Vice-Chancellor of the University of Exeter, agreed with Professor Atkins that school leavers often came to university unprepared, tending to be unable to think critically or independently:

The problem we have with A-levels is that students come very assessment-oriented: they mark-hunt; they are reluctant to take risks; they tend not to take a critical stance; and they tend not to take responsibility for their own learning. But the crucial point is the independent thinking. It is common in our institution that students go to the lecture tutor and say, “What is the right answer?” That is creating quite a gap between how they come to us with A-levels and what is needed at university.²¹⁹

130. We received substantial evidence that teaching to the test, to an extent which narrows the curriculum and puts sustained learning at risk, is widespread. Whilst the Government has allocated resources to tackle this phenomenon and improve practice they fail to accept the extent to which teaching to the test exists and the damage it can do to a child’s learning. We have no doubt that teachers generally have the very best intentions in terms of providing the best education they can for their pupils. However, the way that many teachers have responded to the Government’s approach to accountability has meant that test results are pursued at the expense of a rounded education for children.

131. We believe that teaching to the test and this inappropriate focus on test results may leave young people unprepared for higher education and employment. We recommend that the Government reconsiders the evidence on teaching to the test and that it commissions systematic and wide-ranging research to discover the nature and full extent of the problem.

Narrowing of the curriculum

132. The phenomenon described as ‘narrowing of the curriculum’ is strongly related to teaching to the test and many of the same arguments apply. There are essentially two elements to this concept. First, there is evidence that the overall curriculum is narrowed so

218 Q268

219 Q270

that the majority of time and resources is directed at those subjects which will be tested and other subjects in the broader curriculum, such as sport, art and music, are neglected.²²⁰ Second, within those subjects which are tested, the taught curriculum is narrowed to focus on those areas which are most likely to be tested ('narrow learning') and on the manner in which a component of the curriculum is likely to be tested ('shallow learning').²²¹

133. Doug French of the University of Hull gave his view of the problem:

[Narrowing of the curriculum] is observed particularly in year 6 in primary schools and years 9 and 11 in secondary schools when national tests are taken. In year 6 far too little time is spent on subjects other than those being tested and too much teaching time is devoted to a narrow focus on practising test questions. In secondary schools each subject has its own time allocation, but a narrow test-oriented focus within each subject is commonplace. At sixth form level, the situation, if anything, is even worse with module assessments twice a year leading to AS level after one year followed by A-level in the second year.²²²

134. The QCA observed that the focus on the core subjects leads to relative neglect of the full range of the national curriculum. 90% of primary and 79% of secondary schools reported to the QCA that national testing has led to pupils being offered a narrower curriculum.²²³ Dr Ken Boston also told us that "all the evidence that I hear in my position is about the narrowing of the curriculum that results from these tests".²²⁴

135. The Government, however, states that it makes "no apology for the focus on the core subjects of English, maths and science" as mastery of these disciplines is the key to future success. Pupils who arrive in secondary school without a secure grasp of these subjects to a Level 4 or better will be hampered in their learning of these and other subjects at the higher levels.²²⁵ Sir Michael Barber echoed this view when he gave evidence to us, as did Sue Hackman.²²⁶ The view of the DfES was that:

There is nothing that narrows a pupil's experience of the curriculum so quickly as a poor preparation for the level of literacy and numeracy that the subject demands.²²⁷

136. Whilst the DfES evidence is common sense, it does not really address the concerns raised by other witnesses, including the findings of the QCA reported above. It is true that mastery of the core subjects is vital but, as a consequence of the evidence we cite above that mastery of the examination is given priority over mastery of the subject and that time taken

220 Ev 59; Ev 270; Q46; written evidence from Doug French, University of Hull, para 1.3; written evidence from Association for Achievement and Improvement through Assessment, para 3; written evidence from Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hants; written evidence from Lorraine Smith, Headteacher, Western Church of England Primary School, Winchester, para 7

221 Ev 239; written evidence from Doug French, University of Hull, para 3; written evidence from Advisory Committee on Mathematics Education, para 1; written evidence from Campaign for Science and Education, paras 17-19; written evidence from Association of Science Education, paras 5-6; written evidence from The Mathematical Association

222 Written evidence from Doug French, University of Hull, para 1.3

223 Ev 22

224 Q120

225 Ev 160

226 Q46; Q292

227 Ev 160

to prepare for these examinations is taken from the broader curriculum, children risk missing out on access to a broader range of skills and knowledge. Mick Brookes said that he understood from his colleagues that national testing had narrowed the curriculum and he endorsed a comment attributed to Anthony Seldon, Master of Wellington College, who is quoted as saying:

Children are encouraged to develop an attitude that, if it is not in the exam, it doesn't matter. Intellectual curiosity is stifled and young people's deeper cultural, moral, sporting, social and spiritual faculties are marginalised by a system in which all must come second to delivering improving test and exam numbers.²²⁸

137. The ATL similarly argued that high-stakes testing has had a well-documented narrowing effect on the curriculum and that this has undermined the statutory entitlement of pupils to access to a broad and balanced curriculum, particularly in those schools which fear low scores in the performance tables.²²⁹ Others have deplored focus on core subjects and related tests, leading to the negligence of other subjects of interest to children, and this can be a particular problem in Year 6.²³⁰ The NUT pointed to studies which report that:

[...] high stakes National Curriculum tests had almost wiped out the teaching of some Foundation subjects at Year 6.²³¹

As a result, children whose learning styles do not conform to the content and form of the tests are often missing out on areas of the curriculum in which they may have more success.²³² Simply put, the more creative elements of the curriculum are being displaced by the pressure to teach to the test.²³³

138. The method of assessment has also come in for some criticism. The Association of Colleges gives the example of a written test of mechanical skills or musical understanding which diverts the taught curriculum towards those skills and away from mechanics and music.²³⁴ Thus the clear reliance of national tests on the written, externally marked assessment instrument is contributing to the narrowing of the taught curriculum. Important skills and abilities are ignored because the tests emphasise skills and abilities which are more easily measured.²³⁵ The Association of Colleges calls for a range of assessment methods which would allow for more creativity in the curriculum.²³⁶

139. Given our findings that national tests can only measure a small part of what we might consider valuable in the education of children (Chapter 2) and that teachers are

228 Q128

229 Ev 58

230 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1(f); written evidence from Purbrook Junior School, Waterloo, para 13

231 Ev 263

232 Written evidence from Purbrook Junior School, Waterloo, para 13

233 Written evidence from Doug French, University of Hull, section 1.3; see also a research summary submitted by the Wellcome Trust, Ev 269-270

234 Ev 199

235 Ev 271

236 Ev 199

concentrating their efforts in the classroom on teaching what is likely to be tested, it should come as no particular surprise that many witnesses point to narrowing of the taught curriculum as a particular problem.²³⁷

140. A creative, linked curriculum which addresses the interests, needs and talents of all pupils is the casualty of the narrow focus of teaching which we have identified. Narrowing of the curriculum is problematic in two ways: core subjects are emphasised to the detriment of other, important elements of the broader curriculum; and, for those subjects which are tested in public examinations, the scope and creativity of what is taught is compromised by a focus on the requirements of the test. We are concerned that any efforts the Government makes to introduce more breadth into the school curriculum are likely to be undermined by the enduring imperative for schools, created by the accountability measures, to ensure that their pupils perform well in national tests.

The burden of testing

141. Another theme which manifests strongly in the evidence relates to the quantity of testing²³⁸ and there is concern that the quantity of national testing is displacing real learning and deep understanding of a subject.²³⁹ English school pupils are amongst the most tested in the world.²⁴⁰ Over time, formal national assessment has been applied to ever younger children, so that now even children of four are tested through foundation stage profiling.²⁴¹ Counting foundation stage assessment, a pupil going on to take A-Levels will have been tested in seven of their 13 years of schooling.²⁴² The GTC stated that:

- the average pupil in England will take at least 70 tests during a school career;
- the national testing system employs 54,000 examiners and moderators;
- they deal with 25 million test scripts annually.²⁴³

142. In primary schools, testing takes place through teacher observation at the age of 4 (foundation stage), through moderated teacher assessment at 7 (Key Stage 1) and through formal testing at 11 (Key Stage 2). The NAHT considers unhealthy the dominance of Key Stage tests in primary schools.²⁴⁴ One study estimates that, in Years 5 and 6, the equivalent of three weeks of learning each year is spent on revision and practice tests.²⁴⁵ Sir Michael Barber did not consider that testing in primary schools is overly burdensome over a six-year period.²⁴⁶ Professor Peter Tymms agreed that there was not too much testing *per se*,

237 Ev 70; Ev 269; written evidence from Doug French, University of Hull, section 1.3;

238 Ev 270; written evidence from Doug French, University of Hull, section 1.3

239 Ev 198; Ev 202

240 Ev 46; Ev 51; Ev 73

241 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1(b)

242 Ev 57

243 Ev 74

244 Ev 68

245 Written evidence from Association of Science Education, para 9

246 Q2

but that it was preparation for the tests in a high-stakes context which rendered them problematic.²⁴⁷ Dr Ken Boston of the QCA agreed with this proposition and referred to his concern about the “high stakes put on the assessments because ... they carry 14 different functions”.²⁴⁸ OCR stated its belief that “the sustained, unnecessary and inappropriate mass testing of very young people through the key stage national tests ... is the single biggest cause of the view that there is too much assessment”.²⁴⁹

143. The QCA noted that most primary schools prepare pupils extensively for tests. At Key Stage 2:

- 68% of primary schools employ additional staff;
- 78% set additional homework;
- more than 80% have revision classes and use commercial or QCA practice tests;
- in 80% of primary schools, the amount of time spent on test preparation has increased over the last decade;
- in the second half of the Spring term, 70% of schools spend more than three hours per week on test preparation.

144. The QCA notes a similar pattern of responses from secondary schools.²⁵⁰ In addition, Ofsted reports that schools often deploy their most effective teachers to the end of Key Stage year groups—Years 2, 6 and 9—and teachers in other year groups feel less responsibility for assessing pupils’ progress.²⁵¹ Interestingly, an NUT study found that high-stakes testing causes more concern in the primary sector than in the secondary sector, where long experience of testing and examinations has tended to lead to greater acceptance by teachers and parents.²⁵² By way of comparison, a study for the Royal Society in 2003 found a substantial difference between the time spent in Scottish and English schools on assessment activities at secondary level. English teachers spent more than twice the amount of time each year on assessment activities when compared with their Scottish counterparts at the equivalent of Key Stages 3 and 4. They spent almost seven times more hours at the equivalent of AS/A2 Level.²⁵³

145. In secondary schools, formal national testing takes place at the ages of 14 (Key Stage 3 tests), 16 (GCSEs and equivalents), 17 and 18 (A-levels and equivalents) and witnesses have argued that this is excessive.²⁵⁴ The NAHT is particularly concerned about the dominance of Key Stage 3 tests in secondary schools.²⁵⁵ At the 14–19 stage, it has been argued in

247 Q2

248 Q91

249 Ev 121

250 Ev 32

251 Ev 32

252 Ev 263

253 Written evidence from The Royal Society, section 6

254 Ev 202

255 Ev 68

relation to mathematics, for example, that the prevalence of testing in this age group is having a serious, negative effect on maths teaching, reducing it to “little more than sequences of lessons on test preparation”.²⁵⁶ OCR, on the other hand, argued that assessment spread out over a longer period of time and closer to the learning experience is less stressful than a concentrated period of assessment at the end of a two-year course of study.²⁵⁷

146. In addition, it is currently possible for AS students to sit retakes in order to maximise their grades at the end of the A-level course. It has been argued that this places too great a burden on pupils, diverting them from study of the course to focus on examinations.²⁵⁸ Others, however, argue that retakes have been associated with enhanced understanding of a course for pupils whose marks improved.²⁵⁹

147. Some witnesses have expressed concern over the balance between teacher assessment on the one hand and national testing on the other.²⁶⁰ City and Guilds argued that a considerable burden of assessment is placed on 16–18 year-olds with examinations in each year. This, it argues, could be mitigated if greater use were made of teacher assessment, as is the case with NVQs.²⁶¹ The Chartered Institute of Educational Assessors points to PISA results showing that other countries, such as Finland, achieve good standards with little resort to external assessment and far more emphasis on teacher assessment.²⁶² It has been argued that some national testing should be replaced with moderated teacher assessment or the use of tests drawn from a bank of diagnostic assessments provided centrally by the QCA.²⁶³

148. Contrary to the vast majority of the evidence we have received, the DfES stated in its memorandum that “the statutory assessment system demands relatively little of the child in the eleven years of compulsory schooling”. It is summarised as follows:

- Key Stage 1 tests should be carried out as part of normal lessons and the child will not necessarily notice the difference between the tests and normal classroom tasks.
- Key Stage 2 tests involve one week of testing in May, most tests lasting 45 minutes and the total lasting less than six hours.
- Key Stage 3 tests involve one week of testing, with tests mostly an hour in length and totalling less than eight hours.
- At GCSE, the Government is responding to criticisms and cutting down on coursework.

256 Written evidence from Advisory Committee on Mathematics Education, para 2

257 Ev 121-122

258 Ev 72

259 Written evidence from Mathematics in Engineering and Industry

260 Written evidence from Association of Science Education, para 11

261 Ev 112

262 Ev 222

263 Ev 103; written evidence from Association of Science Education, para 17

- At A-level, the number of units is being reduced from six to four in most subjects.

The Minister told us that no pupil spends more than 0.2% of their time taking tests and stated that “In the end, I flatly reject the argument that there is too much testing”.²⁶⁴

149. We acknowledge the reforms the Government has made to GCSE and A-level examinations. However, the Government must address the concerns expressed by witnesses, among them Dr Ken Boston of the QCA, who see the burden of assessment more in terms of the amount of time and effort spent in preparation for high-stakes tests than in the time taken to sit the tests themselves. This could be achieved by discouraging some of the most inappropriate forms of preparation and reducing the number of occasions on which a child is tested.

Pupil stress and demotivation

150. Many witnesses argued that testing is stressful for children.²⁶⁵ Moreover, repeated testing has a negative effect on children, leading to demotivation, reduced learning potential and lower educational outcomes.²⁶⁶ Testing has even been linked with children’s health, including mental health, problems and lower self-esteem.²⁶⁷ The Association of Colleges, for example, stated that those borderline students who had been assisted with additional resources to get target grades fell victim to false expectations resulting in a sense of inadequacy when they found that they did not have the skills or knowledge to deal with the demands of the next stage of schooling. This, the AoC thought, might account for the high drop-out rate at 17%.²⁶⁸

151. Witnesses have expressed concern that the levels of accountability in schools are resulting in the disillusionment of children.²⁶⁹ Children not reaching the target standard at a given stage have the impression that they have ‘failed’ whilst they may, in fact, have made perfectly acceptable progress.²⁷⁰ Whilst some children undoubtedly find tests interesting, challenging and even enjoyable, others do not do their best under test conditions and become very distressed.²⁷¹ In particular, those children who are not adept at the kind of utilitarian skills and strategies required to do well in tests and who frequently ‘fail’ find the experience “demoralising, reducing self-esteem, including their belief in their ability to succeed with other tasks”.²⁷²

152. Professor Peter Tymms pointed out that this is a complicated area. Children are likely to fail at some things and succeed at others as they go through life. In general, they have a

264 Qq 334 and 368

265 Ev 51; Ev 55; Ev 263; Ev 247; written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 2(a); written evidence from Portway Infant School; written evidence from Purbrook Junior School, Waterlooville, para 8; written evidence from Association of Science Education, para 9

266 Ev 73; Ev 263; Q128; Q172; written evidence from Science Community Partnership Supporting Education, section 2

267 Ev 68; Ev 263

268 Ev 200

269 Q134; written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1(e)

270 Ev 52

271 Ev 272

272 Ev 59

certain natural resilience and expect that, if they try harder, they will do better. However, he considered that it was a mistake to label as ‘failures’ children who did not meet the target standards, even though they may have done very well to get a lower grade or level. He thought it was also an error to label their school as a failure, because children identify with that, too. He concluded that a national monitoring system should examine attitudes, self-esteem, welfare and physical growth, a proposition with which Sir Michael Barber agreed.²⁷³

153. Teaching to the test and narrowing of the curriculum are also thought to have a negative effect on children. The resulting lack of creativity in teaching impacts on children’s enjoyment of a subject and their motivation to learn.²⁷⁴ In the worst cases, teachers may resort to dull and boring methods of teaching, using the looming threat of examinations to motivate pupils rather than inspiring them to learn.²⁷⁵ The Royal Society used the example of science teaching. It argues that the current testing system constrains creativity by giving a high priority to what can be easily measured through written, externally marked examinations. The ability of teachers to meet the individual needs of pupils in the classroom is compromised and, at worst, this may lead to negative attitudes to science amongst pupils, reduced motivation and lower self-esteem.²⁷⁶ An over-emphasis on preparation for national tests at primary level led to a negative effect on children’s enjoyment of science.²⁷⁷

154. The Government has expressed the view that some children will find examinations stressful, but that effective schools will help anxious children to meet the demands which are made of them.²⁷⁸ The Minister told us that he did not accept the idea that the amount of time taken to prepare for national tests was too stressful:

I think that life has its stresses and that it is worth teaching a bit about that in school.

Grade inflation

155. The concept of grade inflation is another phenomenon associated with high-stakes uses of testing. However, before we consider the evidence we have received on this subject, we will clarify the use of some terms. Dr Ken Boston distinguished assessment standards from performance standards. **Assessment standards** denote the degree of difficulty of a test, in his words, the height of the hurdle to be jumped by the student. Through its regulatory role and through its subsidiary, the NAA (National Assessment Agency) which delivers National Curriculum tests, the QCA attempts to maintain the assessment standard constant year on year. **Performance standards** relate to the distribution of students’ grades or levels according to a target standard, in other words, the number of students who clear

273 Qq 27 and 28

274 Ev 270; written evidence from Institute of Physics, para 3

275 Written evidence from The Mathematical Association

276 Written evidence from The Royal Society, para 5; see also written evidence from Association of Science Education, para 2

277 Ev 269

278 Ev 160

the hurdle each year. The QCA's data suggest that the performance standard has been rising.²⁷⁹

156. The concept of grade inflation relates specifically to a reduction in assessment standards, thereby making it easier for students to achieve higher grades or levels. For example, there is an annual debate following publication of GCSE and A-level results as to whether the steady increase in the proportion of students getting the higher grades is genuinely evidence of an improvement in performance standards or whether it is explained by a lowering of assessment standards. Similar arguments have been rehearsed in relation to Key Stage tests.²⁸⁰ However, the debate is really much wider than this because it relates to whether or not national tests are an adequate proxy for the underlying learning and achievement of pupils. To access this debate, we need to consider in more detail the concept of 'standards over time'.

Standards over time

157. The Government notes that the strength and validity of the accountability regime, which is based on performance standards, requires that assessment standards remain consistent over time. The QCA is responsible for this and the Government relates that its processes have been found to be robust and world-class.²⁸¹

158. It is uncontroversial that test scores have improved across the board over time. However, the concept of standards over time is problematic. First, the standards themselves are the result of working groups and various consultations; they embody a series of values, are expressed in everyday language and, as a result of all of this, are open to interpretation by those using them.²⁸² In addition, the descriptions of the standards themselves have changed over time.²⁸³ Professor Colin Richards states that there is no published evidence on the extent to which 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. Without such evidence, it is not possible to be certain that any apparent improvement in performance standards is genuine or an artefact.²⁸⁴

159. Second, the tests have changed over time, some of them radically.²⁸⁵ As Cambridge Assessment put it:

If you want to measure change, don't change the measure. But the nation does—and should—change/update the National Curriculum regularly. Whenever there is change (and sometimes radical overhaul) the maintenance of test standards becomes a particularly aggressive problem.

279 Q55

280 Ev 236

281 Ev 159

282 Written evidence of Professor Colin Richards, Annex A

283 Q27; written evidence of Professor Colin Richards, Annex A

284 Written evidence of Professor Colin Richards, Annex A

285 Ev 226; written evidence from Advisory Committee on Mathematics Education, paras 18-20; written evidence of Professor Colin Richards, Annex A

160. Research suggests that change in the tests does not necessarily mean that there has been a reduction in the assessment standard. It could just be that the things which are measured and the way they are measured are different. It is not a simple matter to establish whether or not an increase in test scores is evidence of a rise in genuine performance standards.²⁸⁶ Even where the curriculum has apparently not changed very much, the way it is taught may have changed considerably.²⁸⁷ In addition, an apparent increase in standards over time according to test scores may be misleading, since the tests arguably measure such a narrow part of the whole curriculum that they are no longer a valid proxy for achievement across the whole of that curriculum.²⁸⁸ The NFER stated that:

There are difficulties in maintaining a constant standard for the award of a level in a high stakes system where tests or questions cannot be repeated. We do though believe that the methods used for this currently which include year-on-year equating and the use of a constant reference point through an unchanging “anchor test” are the best available. A second consideration is that the curriculum coverage each year is limited to the content of that year’s tests.²⁸⁹

161. We are persuaded by the evidence that it is entirely possible to improve test scores through mechanisms such as teaching to the test, narrowing the curriculum and concentrating effort and resources on borderline students. It follows that this apparent improvement may not always be evidence of an underlying enhancement of learning and understanding in pupils.

162. We consider that the measurement of standards across the full curriculum is virtually impossible under the current testing regime because national tests measure only a small sample of pupils’ achievements; and because teaching to the test means that pupils may not retain, or may not even possess in the first place, the skills which are supposedly evidenced by their test results.²⁹⁰

Key Stage tests

163. The ATL highlighted evidence that an apparent improvement in standards of performance has less to do with an improvement in underlying achievement and more to do with familiarity amongst teachers and students with test requirements. It points to research which has demonstrated that changes in tests lead to a sudden fall in performance standards, followed by an improvement as teachers begin to understand how to teach to the new test.²⁹¹ Evidence from the IPPR appears to support this view. It quotes research demonstrating that dramatic improvements in results at Key Stage 2 are not borne out

286 Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing: preliminary evidence about generalization across tests*. Paper presented at the Annual meetings of the American Educational Research Association and the National Council on Measurement in Education held at Chicago, IL.
Cannell, J. J. (1988). Nationally normed elementary achievement testing in America’s public schools: how all fifty states are above the national average. *Educational Measurement: Issues and Practice*, 7(2), 5-9.

287 Ev 60

288 Ev 60

289 Ev 251

290 Ev 60

291 Ev 61

when independent measures are used, which show a much less marked improvement than is suggested by the Key Stage 2 test results.²⁹² On the other hand, the IPPR finds that research evidence on changes in *assessment* standard at Key Stage 2 over time are inconclusive.²⁹³ The IPPR believes that there has been real progress in each of the three core subjects, but less than is indicated by Key Stage test results. It does not consider that the tests have become systematically easier, but thinks that teaching and learning has focused increasingly more narrowly on achieving good test results.²⁹⁴ Professor Peter Tymms broadly agrees with this assessment. His research has led him to the conclusion that the substantial improvements suggested by the test scores were illusory and that there had been some improvement in the underlying attainment in mathematics and writing, but no discernable improvement in reading.²⁹⁵

164. Professor Colin Richards submitted a review of a considerable amount of literature relating to performance standards in primary schools. He summarises his findings as follows:

- The data on performance relate only to three subjects (English, mathematics and science) and only to pupils aged 7 and 11.
- Key Stage test results 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.²⁹⁶

Professor Richards concludes that it is not possible to answer with precision whether standards in primary schools are improving. As other witnesses have suggested, he finds evidence to indicate that there was some rise in performance in the core subjects between 1995 and 2001, but not as great as national test data have suggested.²⁹⁷

292 Ev 236

293 Ev 236

294 Ev 236

295 Q6

296 Written evidence of Professor Colin Richards, Annex A

297 Written evidence of Professor Colin Richards, Annex A

14–19 qualifications

165. In relation to public examinations, the Government points to evidence that A-level standards have remained consistent for at least 20 years, although an increased “breadth of coverage led to a reduced emphasis on some topics”.²⁹⁸ The Government points to other evidence suggesting that A-levels are the most tightly and carefully managed tests at school or any other level; that strategies for maintaining comparable standards across Awarding Bodies are adequate; that Awarding bodies have “broadly consistent and well-regulated systems for setting question papers, managing marking and awarding grades”; and that the QCA has robust systems for monitoring and regulating the work of the Awarding Bodies.²⁹⁹

166. It is highly questionable whether a claim can validly be made that A-levels have remained at a consistent standard over a period as long as 20 years, or indeed anything like it. The DfES itself gave an account of the considerable changes which have been made to this qualification over the years, not the least of which is Curriculum 2000 and piloting of tougher questions in A-level papers to stretch candidates and aid differentiation for universities.³⁰⁰ According to the DfES, the standard required to achieve an A grade will remain the same, but stronger candidates will be able to demonstrate attainment meriting a new A* grade.³⁰¹ The DfES memorandum states:

As our response to criticisms about GCSE and A-level assessment shows, the system has constantly evolved to meet changing needs and it will continue to do so.³⁰²

Without providing some evidence in relation to the underlying assessment standards and the levels against which they are referenced, the Government cannot have it both ways: either standards have been constant over time, or change has been implemented in response to perceived shortcomings in the system. As Edexcel argued:

The curriculum has changed over time, new elements have been introduced and different approaches rewarded. To accurately measure such progress, the curriculum would need to be stable and the same test used each year.³⁰³

167. Research suggests that A-levels have not necessarily become easier, but the examination no longer measures what it used to. Examinations have become more predictable so that teachers have become more effective at coaching pupils and, correspondingly, pupils have become deskilled. It follows, he argues, that one can no longer infer from a top grade that the pupil achieving it has the same skills that a pupil achieving a top grade years ago had.³⁰⁴

298 Ev 159

299 Ev 159

300 Ev 161; Ev 162

301 Ev 162

302 Ev 161

303 Ev 114

304 William, D., Brown, M., Kerslake, D., Martin, S., & Neill, H. (1999). The transition from GCSE to A-level in mathematics: a preliminary study. *Advances in Mathematics Education*, **1**, 41-56.

168. Professor Peter Tymms told us that his research suggested that assessment standards at GCSE appear to be relatively stable over a number of years. However, at A-level they have not, with pupils of a particular ability getting higher grades now than they would have done some years ago. The biggest change has been in mathematics, in which a D grade some years ago would now be the equivalent of a B grade.³⁰⁵ Again, whereas 30% of A-level candidates used to fail, getting less than an E grade, now just a small percentage fail outright, which Professor Tymms characterises as a “dramatic shift”.³⁰⁶ However, an improvement in overall grades does not necessarily mean that assessment standards have been debased. Professor Tymms’ evidence certainly shows that children of the same ability are now getting higher grades than they would have received some years ago, but the reasons for this may be complex. It could, for example, be partially attributed to teaching standards, but the question cannot be answered without an appreciation of equivalent assessment standards.³⁰⁷ Unfortunately, unlike Key Stage tests, A-levels and GCSEs are not pre-tested by the regulator or Awarding Bodies and no items used in previous years are repeated in following years, so there is no way to reference assessment standards from one year to the next. Standard-setting is done retrospectively on the basis of statistical relationships and judgments.³⁰⁸

169. Some witnesses have queried whether a system which allows the existence of multiple Awarding Bodies can ever really ensure that the standards of the assessments produced by those Awarding Bodies are the same.³⁰⁹ We have received evidence that some schools choose a syllabus from a given exam board on the basis that they consider it easier, and therefore more likely that their pupils will achieve higher grades.³¹⁰ In addition, there is some suggestion that some subjects, such as mathematics and the sciences, are ‘harder’ than others, so that pupils and schools are more likely to choose ‘easier’ subjects in an effort to maximise grades.³¹¹

International evidence

170. Witnesses have discussed some of the available international evidence of performance at the secondary stage of education. The essential paradox appears to be that, whilst test scores are improving at home, international rankings are either static or falling. The percentages of pupils achieving the target standards at Key Stages 3 and 4 have risen over time according to domestic test results, yet progress on international attainment measures has stalled. Evidence from TIMSS for Key Stage 3 shows no significant change in performance between 1995 and 2003; and PISA shows that, for a given score at Key Stage 3

305 Q33

306 Q34. This “dramatic shift” may be magnified by the move from a norm referenced to a criterion referenced grading system for A-levels in 1987. Norm referencing means that a pre-set percentage of candidates is awarded each grade. Criterion referencing sets standards against declared criteria of performance, so that the number of candidates achieving a grade A, for example, is not limited by predetermined quotas. Our predecessors reported on this subject: HC 153, House of Commons Education and Skills Committee, *A Level Standards*, Third Report of Session 2002-03, paras 4-8.

307 Q38

308 Qq40 and 45

309 Written evidence from Institute of Physics, 19:7; written evidence from Campaign for Science and Engineering, paras 22-23; The Mathematical Association, Background Paper

310 Q131; Q132

311 Written evidence from Institute of Physics, 19:9; written evidence from Campaign for Science and Education, 21:3

or 4, pupils attained on average a higher PISA score in 2000 than in 2003. Although the UK's response rate to the 2003 PISA survey was too low to ensure statistical comparability, the mean score produced was lower than that in 2000 and led to the UK falling down the international ranking.³¹² One possible explanation, according to the IPPR, is that Key Stage 3 and 4 test scores are not consistent over time. Its preferred explanation, however, is that "improvements in the key stage results do not accurately mirror improvements in underlying pupil attainment, and that some of the improvement is due to more narrowly focused teaching".³¹³

171. It is not possible for us to come to a definitive view on grade inflation in the context of such a wide-ranging inquiry. However, it seems clear to us from the evidence that we have received that the Government has not engaged with the complexity of the technical arguments about grade inflation and standards over time. We recommend that the Government addresses these issues head-on, starting with a mandate to the QCA or the proposed new regulator to undertake a full review of assessment standards.

Accountability through sampling

172. Whilst the use of saturation testing, that is, the testing of each child in a given cohort, is generally agreed to be an appropriate means of ascertaining and certifying individual pupil and, to a certain extent, school achievement³¹⁴, there is rather more argument about whether saturation testing is an appropriate method of testing local and national performance and monitoring the effects of changes in policy.³¹⁵

173. Witnesses have argued for the decoupling of measures of pupil attainment from accountability and monitoring measures in order to remove the need for central collection of individual pupil performance data, thereby removing the high-stakes for the school.³¹⁶ Implicit in this argument is the hope that, once the stakes are removed, the school can get on with the business of teaching children a full and rounded curriculum without fear of recrimination and the children will benefit from the education to which they should, in any event, be entitled.³¹⁷ That the tests would, presumably, remain high-stakes for the individual child has largely been ignored in the evidence we have received.

174. Nevertheless, decoupling accountability and monitoring from a testing system which is primarily designed to measure pupil attainment may have a number of desirable consequences in relation to the issues discussed in this chapter. To summarise the arguments which have been put to us, the incentives for schools and teachers to teach to the test would be reduced considerably.³¹⁸ Likewise, schools and teachers may be more inclined to withdraw from the disproportionate focus on the core subjects of English, mathematics and science, important as these are, and give some more attention to other

312 Ev 58

313 Ev 237-238

314 Written evidence from Doug French, University of Hull, para 1.1

315 Ev 48; written evidence from Doug French, University of Hull, para 1.1

316 Ev 58; Ev 79-81; Ev 265-266

317 Written evidence from Heading for Inclusion, Alliance for Inclusive Education, para 1

318 Ev 258-259

subjects, replacing some of the lost variety in the curriculum. Within the core subjects, teachers may feel more at liberty to take a more creative approach to their teaching which may enhance the enjoyment, satisfaction and even attainment of their pupils.³¹⁹ Less time spent on test preparation would reduce the perception of the testing system as burdensome and, perhaps, result in reduced stress and demotivation for pupils. Finally, there would be scope for developing a system of accountability which is fairer to schools, teachers and pupils alike and which can give some reassurance to the public about the maintenance of assessment and performance standards over time.³²⁰ The IPPR warned, however, that as long as individual pupils sit national, summative tests (albeit separated from the accountability system), that data exists and can be compiled and presented in school performance tables, whether or not the government chooses to collate and publish those tables centrally. Much the same data would be available as before.³²¹

175. It has been widely argued that national cohort sample testing would be a less onerous and more appropriate means of testing local and national performance and monitoring the effects of changes in policy.³²² However, sample testing would not necessarily yield the type of data currently used for individual school accountability. Presumably, if accountability is to be decoupled from national tests designed to measure pupil attainment, different tests or inspections will be required or the concept of school accountability radically overhauled.

176. Dr Ken Boston said the QCA had given advice to the Government on sample testing, but that the Government was more inclined to go in the direction of single-level tests (as to which, see paragraphs 188-198 below), instead setting great store by international sample tests such as PIRLS (Progress in International Reading Literacy Study), PISA (Programme for International Student Assessments) and TIMSS (Trends in International Mathematics and Science Study).³²³ He related that he had told the Government that:

[...] there are many purposes that would be served better by different sorts of tests. Indeed, as you know, some time ago I raised the issue of sample testing, on which the Government were not keen for other reasons.³²⁴

He considered that sample testing, using a standardised test instrument, was the best way of meeting the purpose of discovering national trends in children's performance standards over time. If, on the other hand, the purpose was to compare the performance of school against school, a sample test would not yield the necessary data, but a full cohort test would.³²⁵ He did not believe that Key Stage tests, single-level tests and cohort sampling

319 Written evidence from Association of Science Education, para 18

320 Ev 58; Ev 80; Ev 240

321 Ev 240

322 Ev 48; Ev 55; Ev 74; Ev 103; Ev 116; Ev 230; Ev 240; Ev 259; Q3; Q12; Q138; Q152; Q162; written evidence from Doug French, University of Hull, para 1.1; written evidence from Dr A Gardiner, University of Birmingham, para 9; written evidence from Science Community Partnership Supporting Education, section 3

323 Q88

324 Q93

325 Q100

should be seen as mutually exclusive alternatives: different tests are needed to serve different purposes.³²⁶

177. The Minister, however, did not agree that alternatives to the current Key Stage tests were workable in practice. He acknowledged that some had argued in favour of sample testing to monitor national performance, but thought that testing should also be able to demonstrate a child's progress against a national comparator, as well as measuring the performance of a particular school. He thought that the use of teacher assessment for these purposes was problematic due to the difficulty of assuring comparability of data. He concluded that:

When I look at the matter and begin to unravel the alternatives and think about how they would work in practice, I find that the current SATs are much more straightforward—everybody would understand it. They are used for a series of things, and there might be some compromise involved, but the system is straightforward and simple, and it shows what our priorities are and gives us accountability at every level. I do not think that it is a mess at all.³²⁷

178. The methodology of sample testing is well-established and is used, for example, in international comparison studies such as PISA and TIMSS. It was also used in the UK from the mid-1970s and through the 1980s by the Assessment of Performance Unit (“APU”) within the Department for Education. The APU used light sampling of schools and light sampling of pupils within schools.³²⁸ The GTC sets out a number of advantages to this approach, including reduced burden of testing; anonymity of schools and students, ensuring that the tests are low-stakes; wide curriculum coverage; a range of assessment formats can be employed; test items can be repeated over time; the system is relatively inexpensive; it provides good evidence of performance trends; and it is a tried and tested method. Limitations of the approach include the lack of ratings for individual schools; lack of feedback for individual schools; and certain technical complexities leading to difficulty of interpretation of statistical results.³²⁹ The NFER also pointed out some possible drawbacks with a sampling system. Low-stakes assessment may not motivate pupils to try hard and show what they can really do, resulting in a potential underestimate of ability. In addition, there may be practical difficulties with a system relying on voluntary participation of schools and pupils. However, the NFER broadly supports a regular national monitoring programme.³³⁰

179. The AQA stated that:

[...] a light sampling survey method would enable de-coupling of national assessment from a requirement to deliver robust information on national educational standards. This would enable testing to reflect curriculum change with

326 Q109

327 Q329

328 Ev 84-86

329 Ev 84-86

330 Ev 259

precision, to optimize the learning-focused functions of testing, and enable constant innovation in the form of tests to optimize accessibility.³³¹

Some witnesses have been specific about what they would like to see. The GTC, for example, advocates cohort sampling involving a limited number of pupils in a limited number of schools, using a matrix test structure to allow for multiple tests across the sample to widen the breadth of the curriculum that is being tested. Common questions in any two or more tests would allow for pupils taking different tests to be compared on a common scale. The tests would be administered by teachers, with external support where necessary.³³² The NFER proposed a similar, matrix design.³³³

180. In this context, restoration of the former APU, or something like it, has been a popular theme in evidence.³³⁴ Cambridge Assessment has, however, pointed out a series of technical and political issues which led to the demise of the original APU, stating that its operation was fraught with difficulty. Whilst Cambridge Assessment is in favour of the development of a light sampling, matrix-based model for national monitoring of standards over time, it counsels that this should be done with close attention to the lessons learned from the former APU and from similar systems used internationally.³³⁵

181. We do not necessarily see the point in creating a new body (or reinstating an old one) for its own sake, but we do think that the body developing and administering sample testing for national monitoring should be independent from government and, for this reason, the proposed new development agency, for example, would not be appropriate for this task.³³⁶ As Professor Colin Richards said:

An independent body is needed to keep standards under review and to devise a system for assessing performance in relation to [...] standards over time—at a national level, not at the level of the individual school.³³⁷

182. In summary, the discussion in this Chapter has demonstrated that high-stakes testing, that is, testing where the stakes are high for schools and teachers, can lead to distortion of children's education experience where accountability is linked to the same testing system which is designed to measure pupil attainment:

The full value of a creative, linked curriculum which addresses the interests, needs and talents of all pupils is not exploited because many schools seem to be afraid to innovate when test scores might be affected (even if evidence shows they might go up).³³⁸

331 Ev 208

332 Ev 80-81; further detail at Ev 84-86

333 Ev 258-259

334 Ev 48; Ev 266; Ev 267; Q138; written evidence from Doug French, University of Hull, para 1.1

335 Ev 217-218

336 Ev 266; Ev 267

337 Written evidence from Professor Colin Richards, Annex A

338 Written evidence from Association of Science Education, para 18

183. Whilst we do not doubt the Government's intentions when it states that "The National Curriculum sets out a clear, full and statutory entitlement to learning for all pupils, irrespective of background or ability", we are persuaded that in practice many children have not received their entitlement and many witnesses believe that this is due to the demands of national testing.

184. We are persuaded that the current system of national tests should be reformed in order to decouple the multiple purposes of measuring pupil attainment, school and teacher accountability and national monitoring. The negative impacts of national testing arise more from the targets that schools are expected to achieve and schools' responses to them than from the tests themselves.

185. School accountability should be separated from this system of pupil testing, and we recommend that the Government consult widely on methods of assuring school accountability which do not impact on the right of children to a balanced education.

186. We recommend that the purpose of national monitoring of the education system, particularly for policy formation, is best served by sample testing to measure standards over time and that cohort testing is neither appropriate nor, in our view, desirable for this purpose. We recommend further that, in the interests of public confidence, such sample testing should be carried out by a body at arms length from the Government and suggest that it is a task either for the new regulator or a body answerable to it.

5 Reform

187. In this Chapter, we look at three areas of reform which have featured as important in this inquiry: the pilot study of single-level tests; the new Diploma qualification; and the proposals for a new regulator and test development agency. This should not be taken as a full review of each of each area of reform, rather as an indication of some of the testing-related issues which we believe are likely to become important as these initiatives are developed further.

Single-level tests: the Making Good Progress pilot

188. The Government introduced pilot tests, resulting from the Making Good Progress consultation document, in around 500 schools from September 2007. Although the pilot study is not due to finish until Summer 2009, the Government appears to have thrown its weight behind the scheme in the *Children's Plan*, which states that:

It is our intention to implement new single level tests [...] on a national basis at the earliest opportunity, subject to positive evidence from the pilot and to endorsement of this approach from the Regulator.³³⁹

189. The new tests are known as 'single-level tests'. In contrast to the current Key Stage tests, which test students only at the end of a Key Stage and across a range of levels, resulting in a level being "awarded" as a grade, the single-level tests assess the pupil at a set, 'single' level, which the candidate either achieves or does not achieve. Teachers at Key Stage 2 and 3 in English and mathematics will be able to enter a pupil for an externally written and marked test as soon as, in the teacher's assessment, the pupil is ready to be tested. Science will continue to be tested with Key Stage tests while the Government explores "new options" for the assessment of science.³⁴⁰ The principle of 'testing when ready' means that the tests will be available twice a year.

190. A new performance target for schools will be set, with children expected to make two levels of progress between each Key Stage. One-to-one tuition will be offered to pupils making "slow progress".³⁴¹ The single-level tests are intended to be confirmatory of a teacher's assessment of the level of attainment of a pupil. Professor Dylan Wiliam argues that this proposition is disingenuous: "If the teacher's judgment is that a pupil has reached a level, but the test indicates that they have not, there is no process of reconciliation between these two. It is the teacher who is wrong, and the test that is right. The role of the teacher's assessment is therefore limited to deciding when the student should take the test [...]"³⁴²

191. Several concerns with the pilot tests have been raised in evidence to this inquiry. The availability of tests twice each year and the necessity that a child sit a separate test for each

339 *Children's Plan, Cm 7280, para 3.68*

340 *Children's Plan, Cm 7280, para 3.68*

341 *Children's Plan, para 3.61*

342 Making good progress: incorrect diagnosis and weak treatment? Prof Dylan Wiliam, p3

National Curriculum level means that a child potentially faces a much larger number of national tests and certainly more frequent testing than under the current end of Key Stage regime. Some witnesses have highlighted possible curriculum disruption and logistical difficulties with single-level tests if testing is to take place over a longer period of time. There may also be additional costs involved.³⁴³

192. Other witnesses have criticised single-level tests on the basis that the proposals pay no heed to the effects of measurement error. Candidates can keep re-sitting each test until they are successful and pass the next level. As borderline students finally succeed, standards will appear to have risen over time, yet the rise will be attributable only to a statistical artefact.³⁴⁴ Once achieved, pupils are deemed to be at that level in perpetuity. This, it is argued, is not defensible since it is possible for children to fall back in subjects if they are neglected.³⁴⁵ There is suspicion that this ‘one-way ratchet’ mechanism will entrench teaching to the test and narrowing of the taught curriculum as schools will be held accountable on the basis of the number of levels of progress which a child makes across a Key Stage. Professor Dylan Wiliam argues that teaching to the test under the single-level test regime will “permeate the entire key stage, rather than the final year as it does now”.³⁴⁶

193. The National Foundation for Educational Research does not believe that the single-level test model will support teaching in any direct way and further states that:

The achievement of a level and the knowledge that it cannot be removed may act to demotivate rather than motivate. We would advise that the ‘one way ratchet’ is abandoned and that the system allows for re-testing of doubtful cases so that high levels of certainty are achieved and so that misclassification is minimised.³⁴⁷

194. There is some concern about the conduct of the pilot, not least because the schools taking part are not released from their Key Stage testing obligations, which they must continue to run in parallel.³⁴⁸ Some teachers have complained about the burden of the pilot tests and a proportion of schools have already dropped out.³⁴⁹ The pilot study caused further controversy when it was reported that the Government was delaying notification of the results of the first round of single-level tests results. David Bell later stated in evidence that:

we would not and should not be surprised that, when you pilot a new form of testing, you might need to see what actually happened. We are doing some further work, and we have asked the National Assessment Agency to do some further work. We are not ready yet to come back with the results of that analysis or to say what has happened.³⁵⁰

343 Ev 70-71; Ev 50-54; Ev 246; Ev212-214

344 Ev 204

345 Ev 252

346 *Making good progress: incorrect diagnosis and weak treatment?* Prof Dylan Wiliam, p4

347 Ev 252

348 TES, *Schools back off test trial*, 14 December 2007, p1

349 TES, *Schools back off test trial*, 14 December 2007, p1

350 Q311

195. After the results had been released to schools, the Minister explained that there had been some “unexpected patterns” and “unusual outcomes” in the results.³⁵¹ The Minister expressed surprise at the number of candidates had been entered at the wrong level.³⁵² He explained further that:

the most significant unusual outcome was variations between Key Stage 2 and Key Stage 3 pupils taking the same test. So, let us say that they were taking a level 4 writing test. The Key Stage 2 students were doing significantly better when they were taking exactly the same test as the Key Stage 3 students. Now, that was a bit odd.³⁵³

196. The NAA’s subsequent investigation found a number of factors which were likely to have combined to produce these unusual outcomes:

- inappropriate entry of pupils not secure in the level for which their teachers had assessed them;
- a style of test unfamiliar to both pupils and markers, with questions pitched at a single level, rather than a range of three levels;
- less motivated pupils; research suggests that pupil motivation for new tests taken in a pilot may be lower than for National Curriculum statutory tests, and that this factor may be more marked for pupils in Key Stage 3 than in Key Stage 2;
- a number of pupils not completing test papers, particularly on higher level papers; lack of familiarity with this type of test may have contributed to this;
- markers unaccustomed to marking scripts at a single level from pupils in two key stages.

197. The Minister wrote to explain that:

It is important to recognise that NAA developed these tests on a much shorter timescale than is typical for test development, and that this did not allow for the usual pre-testing that would take place.³⁵⁴

However, giving oral evidence, he reminded us that:

We should bear in mind that it took four years for the SATs to be piloted.³⁵⁵

We have not received any evidence to indicate why the Government might be in such a hurry to roll out these single-level tests. When so much is at stake, we consider this haste inappropriate at best.

198. Our predecessors warned the Government about bringing in new tests with undue haste. We recommend that the Government allows sufficient time for a full pilot of the

351 Ev 176-177

352 Q373

353 Q380

354 Ev 177

355 Q380

new single-level tests and ensures that any issues and problems arising out of that pilot are fully addressed before any formal roll-out of the new regime to schools.

Personalised learning and Assessment for Learning

The Government consultation *Making Good Progress* emphasises more informal teacher assessment and personalisation in teaching and the *Children’s Plan* has now reinforced this approach.³⁵⁶ Teachers’ assessment skills are, therefore, ever more important with the increasing focus on personalised learning and monitoring of pupil progress. The Government has made a strong commitment to the techniques embodied in Assessment for Learning (“AFL”)³⁵⁷ and has committed resources for professional development of teachers in AFL techniques.³⁵⁸ The Children’s Plan states that “Our new approach in schools—which looks at progression across stages—means we will focus on every pupil, in every year group, not just those at the end of key stages and in the middle of the ability range.”³⁵⁹ The Government considers that greater personalised learning under the “new approach” will “help to identify and prioritise those pupils who are in danger of stalling in their learning at the start of secondary school”.³⁶⁰

199. The QCA explains that AfL involves using assessment in the classroom to raise pupils’ achievement. It is underpinned by the proposition that pupils will improve most if they understand the aim of their learning, where they are in relation to this aim and how they can achieve the aim or reduce the gap. Effective AfL is already used in classrooms by some teachers. The key characteristics of AfL are:

- teachers using effective questioning techniques;
- teachers using marking and feedback strategies;
- teachers and pupils sharing learning goals; and
- peer and self-assessment by pupils.

These characteristics of AfL are embodied in a number of processes:

- sharing learning goals with pupils;
- helping pupils know and recognise the standards to aim for;
- providing feedback that helps pupils to identify how to improve;
- believing that every pupil can improve in comparison with previous achievements;
- both the teacher and pupils reviewing and reflecting on pupils’ performance and progress;

356 DfES consultative document, 8 January 2007; *Children’s Plan*, p67

357 *Making Good Progress*, p11

358 *Children’s Plan*, para 3.65; Ev 78

359 *Children’s Plan*, para 3.54

360 *Children’s Plan*, para 3.79

- pupils learning self-assessment techniques to discover areas they need to improve;
- recognising that both motivation and self-esteem, crucial for effective learning and progress, can be increased by effective assessment techniques.³⁶¹

The QCA states that research has shown that participation in the review process raises standards and empowers pupils to take action to improve their performance. AfL is a type of formative assessment and is different from assessment of learning, otherwise known as summative assessment, which involves judging pupils' performance against national standards (level descriptions) as is the case with Key Stage tests. However, according to the QCA, the formative use of summative data remains an important aspect of AfL.³⁶²

200. The NAHT states that AfL is vital and goes far beyond snapshot, national tests. Many schools have developed “sophisticated pupil tracking systems” using these methods.³⁶³ The GTC agrees, but admits that “there remains considerable diversity in school approaches to AfL”.³⁶⁴ According to the ASCL:

The use of assessment for learning has improved the quality and extent of formative assessment, encouraging students to think more about their own learning and helping teachers to mould their teaching style more effectively to the needs of the students.”³⁶⁵

201. The ASCL finds the “personalised classroom” an attractive prospect, but argues that it can only be a reality if the teacher has access to necessary and reliable data. The current regime of Key Stage tests, they argue, does not provide such data because of their high-stakes nature and “weakness” of the tests themselves. Other, diagnostic testing mechanisms are needed.³⁶⁶

202. The ASCL and others favour a move towards a stronger element of teacher assessment in the classroom, with teachers able to draw on a national bank of tests developed under the auspices of the QCA and the administration of which would be monitored by chartered assessors.³⁶⁷ However, the ATL argues that there is a need to address perceptions of bias in teacher assessment.³⁶⁸ The NASUWT cautions against approaches to formative assessment which are overly bureaucratic and burdensome for teachers, particularly in terms of the need to demonstrate effective performance to auditors.³⁶⁹

203. Perhaps one way of making sense of the rather diverse evidence we have received on this subject is to adopt the matrix analysis put forward by the NFER. It contrasted the

361 Source: http://www.qca.org.uk/qca_4337.aspx

362 Source: http://www.qca.org.uk/qca_4337.aspx

363 Ev 68-69

364 Ev 78

365 Ev 48

366 Ev 51

367 Ev 266; Ev 49

368 Ev 55-56

369 Ev 248

formative/diagnostic and summative dimensions of assessment with informal and formal processes of assessment and presented the information in Table 3:

Table 3: The four quadrants of educational assessment

	Purposes	
Processes	Formative	Summative
Informal	<i>Questioning</i> <i>Feedback</i> <i>Peer assessment</i> <i>Self assessment</i>	<i>Essays in uncontrolled conditions</i> <i>Portfolios</i> <i>Coursework</i> <i>National Curriculum teacher assessment</i>
Formal	<i>Analysis of tests, exams, essays</i> <i>Target setting</i>	<i>Tests</i> <i>Exams</i> <i>Essays in controlled conditions</i>

Source: NFER

204. According to the NFER, AfL is a means of **informal formative assessment** which can have extremely positive results but may make “formidable demands on teachers in terms of their professional knowledge and skill”.³⁷⁰ The NFER strongly supported the principles of AfL and argued that better support materials should be provided to teachers in order to encourage the spread of formative assessment in classrooms. It cautioned, however, there are limitations to AfL. More research was needed in order to understand the precise aspects of AfL which lead to greater gains in pupils’ knowledge and understanding. In addition, AfL data is not useful for gaining an overview of the overall level of attainment or of the curriculum as a whole, since its focus is on what has just been learned and what is about to be learned. The use of teacher assessment, self-assessment by pupils and peer-assessment by classmates gives rise to problems with reliability and bias in the data. There are additional problems inherent in the systematic collation of these data, which may be time-consuming, so as to allow for reliable and comparable overall judgements. For these reasons, NFER considered that AfL neither could nor should provide summative information: this function should be served by a separate assessment system.³⁷¹

205. **Formal formative assessment** is increasingly a focus of the Key Stage testing system in the sense that test results are systematically analysed (RAISEonline would be an example) in order to generate information for teaching and learning. There are problems with using such information for formative purposes, however. For example, formative information is most useful at the beginning of a programme of study, yet the Key Stage tests are, by definition, at the end of the Key Stage. The NFER suggested that e-assessment

370 Ev 254

371 Ev 254

may provide a useful, unobtrusive method for formal formative assessment and is currently researching in this area.³⁷²

206. **Informal summative assessment** is already an integral part of the testing system in that the Key Stage testing system requires teacher assessment judgements alongside test results. However, the status of these teacher assessments has tended to be lower than that accorded to the test results themselves. The NFER noted that the balance has started to change as the themes of AfL have been integrated into policy and there is renewed interest in systematic, informal summative assessment, exemplified by the QCA's Assessing Pupils' Progress (secondary schools) and Monitoring Children's Progress (primary schools) initiatives. In Wales, the balance has changed more radically, Key Stage tests having been replaced by teacher assessment.

207. The NFER noted that, in order to be used summatively, teacher assessment information must be tied to the standards embodied in the National Curriculum level descriptions. However, these descriptions are broad and include imprecise judgemental terms. A consensus within the teaching profession on their meaning and application is necessary if summative teacher assessments are to be meaningful. This would, in turn, involve an extensive moderation process which would be "professionally valuable but costly and extremely time-consuming".³⁷³ This is part of an ongoing debate about the potential for teacher assessment to replace test results as the main source of formal summative assessment. The NFER pointed out that, on the one hand the scope of assessment and teacher involvement would be enhanced; on the other there were serious questions about manageability and reliability which would need to be addressed. The NFER set out three conditions for its successful introduction: major investment in professional development in relation to the criteria for assessment; professional development to enhance understanding of the nature and purposes of the four quadrants of assessment as set out in Table 3; and a system of external monitoring and accountability to assure public and professional confidence.³⁷⁴

208. Finally, **formal summative assessment** is exemplified by Key Stage tests; the single-level tests are a new evolution in this quadrant of the NFER model. Formal summative assessment may serve many different purposes, as was discussed in Chapter 2, and we have argued that the current Key Stage tests are a compromise which attempts to meet a wide variety of these purposes, including assessing pupil attainment, school accountability and national monitoring. The NFER expressed the view that the Key Stage tests adequately serve the first two of these purposes, but serve less well the purpose of national monitoring.

209. The NFER stated that introduction of single-level tests, if they are to replace the Key Stage tests, should be accompanied by a statement of which purposes they are expected to meet, which they are not, and the extent to which they meet the requirements for validity, reliability and manageability for each of the intended purposes.³⁷⁵ It is at this point that the utility of the NFER's model becomes apparent. It considers that all four of the quadrants of

372 Ev 255

373 Ev 255

374 Ev 255-256

375 Ev 261

the model are essential to the effective education of children. Each has distinctive features and requirements, yet all are related and education professionals and policy-makers alike should accord appropriate attention to each.³⁷⁶ Part of the problem with the single-level tests appears to be a confounding of the features and requirements of the four quadrants of the model. The NFER supported testing when ready and closer ties between tests, teaching and learning and considered that such notions are consistent with personalised learning and AfL. However, it doubted that the single-level tests as described in *Making Good Progress* would promote personalised learning and AfL, as claimed by the Government. The single-level tests would give an indication of pupil attainment, but would not simultaneously provide diagnostic information to indicate appropriate next steps in learning. The tests are likely to be too far apart to be useful in identifying the detail necessary for personalised learning: a level represents, on average, two years of teaching. This means, according to the NFER, that the single-level tests are unlikely to support teaching in any direct way. Looked at from another direction, for the reasons set out in paragraph 204, AfL techniques are unsuitable for use in generating summative data, so the Government's claims in *Making Good Progress* that single-level tests will serve both summative and formative/diagnostic purposes would appear to be wide of the mark.

210. *Making Good Progress* characterises single-level tests as integral to personalised learning and Assessment for Learning yet also the means by which to generate summative data. We agree with the National Foundation for Educational Research that this single assessment instrument cannot validly perform these functions simultaneously and, if it is attempted, there is a danger that the single-level tests will work for neither purpose. The single-level tests may be useful, however, if their purpose is carefully defined and the tests are developed to ensure they are valid and reliable specifically for those purposes.³⁷⁷

211. We recommend that, if single-level tests are introduced, they are used for summative purposes only and that Assessment for Learning and personalised learning are supported separately by enhanced professional development for teachers, backed up with a centralised bank of formative and diagnostic assessment materials on which teachers can draw as necessary on a regular basis.

Progression targets

212. It is a feature of *Making Good Progress*, reiterated in the Children's Plan, that national testing will continue to be used for the purpose of school accountability. In the context of single-level tests, the proposed target is that a child should move up two levels between Key Stages. These will be known as 'progression targets' and the results will be published in performance tables.³⁷⁸ These proposals are unpopular with the many organisations who have submitted evidence taking issue with the current targets and tables. Although the progression target is that pupils will progress two levels between Key Stages, the Key Stages are not of equal length, and the progression target has been criticised as arbitrary and

376 Ev 258

377 Ev 261

378 *Children's Plan*, para 3.67

unfair on this basis.³⁷⁹ In addition, witnesses have objected that progression targets assume that all pupils should ideally progress at the same rate, which is considered by many as a false assumption.³⁸⁰

213. Witnesses have pointed to a contradiction between the concept of personalised learning, which recognises differences in the abilities and needs of children, and systemic targets, which assume that children should ideally develop at the same rate, that is, two levels across each Key Stage.³⁸¹ The NUT warned that, in its view, the proposed progression targets linked to the single-level tests will ultimately perpetuate the current practice of diverting “resources towards children on the borderline of national target levels”.³⁸² The NASUWT maintains that many of the problems associated with testing relate to the high-stakes environment in which it takes place and argues that associating single-level tests with accountability measures based on progression of pupils leads to a “significant danger that such an approach would result only in the replacement of one high-stakes assessment system with another”.³⁸³ The ATL similarly believes that its vision for reform, which includes the use of AfL, personalised learning and teacher assessment, cannot exist alongside league tables “which already have a pernicious effect on the current national testing system.”³⁸⁴

214. The NAHT sees single-level tests as providing data of significant value to a school. However, if those data are used in the same way as they are currently for school accountability, there is no reason to assume that the new data set derived from single-level tests “would be any more accurate or less damaging than the current data set” if it is to be used in the same way and in isolation from other measures of performance.³⁸⁵ Given the emphasis on personalised learning, according to the NAHT, data from single-level tests will not support comparisons of performance between different schools.³⁸⁶

215. Single-level tests may have some positive effects and we certainly approve of the Government’s new emphasis on the personalised approach. However, the Government has structured the single-level testing system in such a way as to risk a transposition of existing, systemic problems into the new arrangements. Without structural modification, we foresee that the existing problems—including teaching to the test, narrowing of the taught curriculum and the focus on borderline candidates to the detriment of others—will continue under the single-level test regime.

216. We believe that true personalised learning is incompatible with a high-stakes single-level test which focuses on academic learning and does not assess a range of other skills which children might possess. Children who struggle with the core subjects may receive more targeted assistance in those subjects. However, if this means that children

379 Ev 52-53; Ev 248; written evidence from the Advisory Committee on Mathematics Education;

380 Ev 247-248

381 Ev 51

382 Ev 264

383 Ev 246-247

384 Ev 56

385 Ev 70-71

386 Ev 70

who are struggling with core subjects get less opportunity to access the wider curriculum, they risk being put off learning at an early age. We call upon the Government to invest in ways to help and, if necessary, train teachers to improve the basic skills of struggling pupils while enhancing their enjoyment of learning and guaranteeing their access to a broad curriculum.

217. We are concerned about the “one-way ratchet” on the attainment of test levels under the single-level testing regime and we find persuasive the evidence that this may lead to an apparent, but artificial, improvement in performance standards. We recommend that the Government consider further whether it is in children’s best interests that they may be certified to have achieved a level of knowledge and understanding which they do not, in truth, possess. We suspect that this may lead to further disillusionment and children perceiving themselves as ‘failures’.

218. We recommend that the Government urgently rethinks its decision to use progression targets, based on pupils’ achievement in single-level tests, for the purposes of school accountability. If such high-stakes accountability measures are combined with more frequent testing of children, the negative effect on children’s education experiences promises to be greater than it is at present. We urge the Government to listen to the QCA, which has already warned of the dangers of saddling the single-level tests with the same range of purposes which the Key Stage tests demonstrably cannot bear.

Diplomas

219. The Diploma is a new, employer-designed 14–19 qualification, designed to give the student a rounded education, combining theoretical and practical learning. It combines essential skills and knowledge, practical experience, employer-based learning, as well as functional English, mathematics and ICT and the opportunity to develop a specialism or complementary study. The first five Diplomas—Construction and the Built Environment; Creative and Media; Society, Health and Development; Information Technology; and Engineering—are available in 2008 in selected areas. Others will be added in future years, with three new Diplomas in Science, Humanities and Languages starting in 2011, leading up to an entitlement to 17 Diplomas for 16–18 year-old students from 2013.³⁸⁷ The three principal components, together with their characteristics are set out in Figure 2.

387 Ev 165; DCSF, *Promoting achievement, valuing success: a strategy for 14–19 qualifications*, Cm7354, March 2008

Figure 2: principle components and characteristics of Diplomas

Components	Characteristics
Principal learning	<p>Gives the industry title of the Diploma</p> <p>Learning that is related to the sector of the economy</p> <p>Learning that is designed and endorsed by industry</p>
Core content	<p>Includes the assessment of Functional Skills in English, mathematics and ICT</p> <p>Develops a student's employability skills of teamwork and self management</p> <p>Gives the student the opportunity to produce an extended project</p> <p>Requires at least 10 days' compulsory work experience</p>
Additional and/or specialist learning	<p>Allows for the student to specialise</p> <p>Allows for the student to choose more qualifications</p> <p>Allows for flexibility and choice of learning</p>

Source: QCA, 14-19 education and skills: what is a Diploma?³⁸⁸

220. The QCA has also produced a diagram setting out how the 14–19 qualifications fit together. This is reproduced at Figure 3.

Figure 3: 11–19 progression routes

19+	Employment		Higher education		Further education
16–19	Other work-based learning	Advanced Apprenticeships; Apprenticeships	Advanced Diploma Foundation and Higher Diplomas also available	A -levels International Baccalaureate	
Functional skills in all learning routes					
14–16	Young Apprenticeships	Higher Diploma	GCSEs	Foundation learning tier	Foundation Diploma
11–14	Key Stage 3				

Source: QCA, 14-19 education and skills: what is a Diploma?³⁸⁹

221. The introduction of the Diploma has taken some time and the shape of the qualification has been through several evolutions. The report of the Working Group on 14-19 Reform, established in February 2003 and chaired by Mike Tomlinson, proposed that all education and training for 14–19 year-olds should be brought together in a common format of learning programmes together with a unified system of certifying achievement in those programmes.³⁹⁰ Our predecessors expressed disappointment that the Government decided not to implement in full the proposals of the Working Group and create a unified, overarching Diploma to replace the current qualifications system.³⁹¹ Instead, the Diploma emerged as a further qualification alongside existing qualifications. However, as the framework has continued to evolve, we have learned that Advanced Diplomas will be the equivalent of three and a half A-levels and that the Government intends to “bring the best of existing qualifications within the Diploma framework”.³⁹² With a full Government review of Diplomas, GCSEs, A-levels and other general qualifications announced for 2013, we are beginning to suspect that the wheel may have turned full circle and that the Government intends to adopt the Tomlinson proposals after all.

222. Greg Watson of OCR labelled the Diploma “the most complicated qualification that I have ever seen” and emphasised the urgent need to recognise its complexity and address the practical logistics of how it will be taught.³⁹³ As well as studying for a wide variety of different components of the course, students will have the opportunity to work in different schools, colleges and work places. Indeed the QCA states that no one school or college will be able to teach the entire range of available Diplomas.³⁹⁴ Schools and colleges will be required to work in collaboration with each other, and alongside work-based learning providers, and some witnesses have expressed concern that the current accountability regime, which puts schools in direct competition with each other, is incompatible with this aim.³⁹⁵ The GTC, for example, considers that the current accountability structures do not sit easily with a cross-institutional, collaborative approach amongst schools and colleges and views the introduction of Diplomas as an opportunity to move away from:

[...] an assessment system dominated by the purposes of quality control and accountability and assessment of learning towards a more balanced model with a greater element of diagnostic and formative assessment for learning.³⁹⁶

223. Witnesses have generally welcomed the Diploma.³⁹⁷ Jerry Jarvis of Edexcel thought that the Diploma’s broad curriculum might increase the breadth of experience and learning of those entering higher education.³⁹⁸ Some have highlighted the positive benefits

389 http://www.qca.org.uk/libraryAssets/media/QCA_Diplomas_6pp.04.pdf

390 *14–19 Curriculum and Qualifications Reform: The final report of the Working Group on 14–19 Reform*, October 2004

391 HC 246, House of Commons Education and Skills Committee, *14–19 Diplomas*, Fifth Report of Session 2006–07, paras 13–15

392 DCSF, *Promoting achievement, valuing success: a strategy for 14–19 qualifications*, Cm7354, March 2008, p21

393 Q197

394 QCA, *14–19 education and skills: what is a Diploma?*
http://www.qca.org.uk/libraryAssets/media/QCA_Diplomas_6pp.04.pdf

395 Ev 198; Ev 201

396 Ev 76

397 Ev 72–73; Ev 143; Ev 230; Q265; Q275; Q286

398 Q195

of a balance between internal and external assessment.³⁹⁹ City and Guilds, however, pointed to a tension between the general and vocational themes of the Diploma which, in its view, will be played out in the chosen methods of assessment. The vocational theme would indicate an emphasis on performance evidence, whereas the general theme would indicate an emphasis on knowledge-based evidence. City and Guilds considers it too early to say in what form the Diploma will finally emerge.⁴⁰⁰ The ASCL is also sceptical, stating that:

Experience of previous attempts to introduce quasi-vocational qualifications, for example GNVQ, lead ASCL members to be concerned that the assessment of the diplomas may be too much like those of GCSE and A-level. Effective vocationally-oriented courses cannot be assessed in the same way as academic courses.

224. However, the DfES memorandum referred to the need for “innovative forms of assessment to reflect the blend of practical and theoretical learning” and stated that assessment would combine locally determined and standardised external assessment that would provide both formative and summative data on students’ progress (and, crucially, the performance of educational institutions).⁴⁰¹ In addition, the DCSF confirmed that Diplomas will combine “internal controlled assessment” with a practical focus, with theory-focused external assessment.⁴⁰²

225. We welcome the Government’s stated intentions that both the vocational and the general elements of Diplomas should be reflected in the methods of assessment used. We caution the Government against any haste in shifting this delicate balance in future until the full implications of such a shift have been understood.

226. The Chartered Institute of Educational Assessors and others warned of the vulnerability of a new qualification. Steps must be taken to ensure that the Diploma is wanted by students, parents and users of qualifications, such as employers and higher education.⁴⁰³ Jerry Jarvis of Edexcel has warned that “if the diploma doesn’t earn its spurs as a qualification, and that means respect from employers, pupils, parents and higher education, we face a serious problem. There is a huge educational risk to this country.”⁴⁰⁴ We are concerned that in a recent survey conducted by ACS International Schools, fewer than 4 in 10 university admissions officers saw the Diploma as a “good alternative to A-Levels”.⁴⁰⁵ In fact, the CIEA finds “evidence of a real intention to make the new qualification work and of cooperation across educationalists, employers and awarding bodies”.⁴⁰⁶ Jerry Jarvis told us that Edexcel and the other Awarding Bodies were certainly extremely keen to make the Diploma work because of the “huge investment” that they have

399 Ev 73; Ev222; Ev 224

400 Ev 112; Q198

401 Ev 162

402 Ev 165

403 Ev 230; see also Jerry Jarvis, Q196

404 http://dailymail.co.uk/pages/text/print.html?in_article_id=560186&in_page_id=

405 <http://news.bbc.co.uk/1/hi/education/6917842.stm>

406 Ev 230; See also Prof Steve Smith, Q276

made in it.⁴⁰⁷ Greg Watson of OCR added that there would inevitably be issues which needed to be addressed over time, but the fate of Diplomas was likely to be decided by the way they are taught in the first few years. For that reason, it was imperative that teachers received all the support that Government and the Awarding Bodies could give them.⁴⁰⁸

227. We are concerned that the results of a recent NUT survey showed that, within the schools introducing the Diploma this year, the majority of staff are still unfamiliar with them.⁴⁰⁹ Jerry Jarvis of Edexcel has expressed concern that teachers will receive only three days' training before the roll-out in September.⁴¹⁰ The NAHT has also expressed concern about the lack of training provided to teachers.⁴¹¹

228. Whilst welcoming the Diploma, the NAHT remains sceptical as to whether the opportunity for a radical and imaginative approach to assessment is actually taken.⁴¹² However, the NAHT considers that:

If anything will assist the reintegration of some of the NEETs (young people not in education, employment or training) it will be the further, suitable development of modular, component assessment within the new vocational diplomas.⁴¹³

According to OCR, a suspicion of alternative qualifications (ie other than A-level) which assess the practical application of skills may reflect the belief, in turn reflected in Government policy, that the only rigorous way to assess achievement is through formal, written examination.⁴¹⁴ OCR states that, in its experience, new qualifications take at least ten years to become accepted and take root.⁴¹⁵ On this basis, there is plenty of time for Diplomas to be altered radically from their current format. OCR has already noted that:

In seeking parity with GCSE and GCE, the main parts of the Diplomas have increasingly adopted models which mirror the models for GCSE/GCE laid out in the regulatory codes of practice. The grading structures have also been adopted to mirror GCSE/GCE scales.⁴¹⁶

However, Jon Coles, Director of 14–19 Reform at the DCSF, reassured us that Diplomas were intended to introduce a broader range of assessment methods to test the broader range of skills which had been called for by universities and employers.⁴¹⁷ Furthermore, the Minister told us that “The fundamental design of the Diplomas will not change”. What he terms ‘generic learning’—“functional skills, personal learning and thinking skills”—will

407 Q196; Q200

408 Q200

409 http://www.dailymail.co.uk/pages/live/articles/news/news.html?in_article_id=560186&in_page_id=1770

410 Ibid.

411 <http://www.telegraph.co.uk/news/1922616/New-diplomas-may-fail,-warn-headteachers.html>

412 Ev 72

413 Ev 73

414 TA46, para 27

415 TA46, para 4

416 TA46, para 34

417 Q322

remain part of the Diploma curriculum. He considers that Diplomas have had a fair lead-in time, “It is not the full OCR 10 years, but it is fair.”⁴¹⁸

The extended project

229. The extended project forms part of the core content of a Diploma and may be used as a free-standing qualification which can be taken alongside A-levels, in which case it is expected that an extended project would be taken instead of, not in addition to, a fourth or fifth AS Level. The project will be in an area of the student’s choice, to be approved by the relevant Awarding Body, and will test skills such as independent research, study and planning.⁴¹⁹

230. Some witnesses have remarked on the apparent paradox that, as coursework is being scaled back considerably for GCSE and A-level, the extended project is being introduced for Diplomas (and also for some A-level courses).⁴²⁰ The DCSF addressed this point in a supplementary memorandum, stating that the theoretical focus of GCSEs and A-levels makes external assessment more appropriate than coursework. Diplomas, on the other hand, focus more on practical learning, making coursework more appropriate. The Department has recharacterised coursework as “internal controlled assessment” for GCSEs, A-levels and Diplomas.

231. The NASUWT expressed concern that the extended project may prove to be overly burdensome for teachers, who must ensure that a student’s learning is assessed validly, particularly in respect of reliability and comparability. Although it acknowledges extended projects as beneficial for students’ learning, the NASUWT cautions that this benefit may be undermined by bureaucratic and work-intensive procedures for assessing the extended project.⁴²¹

232. Although the reaction to the introduction of Diplomas has been one of cautious welcome, there are important caveats. The issue of accountability has arisen once again, and we consider that it is of the first importance that the Government addresses this issue once and for all. There is concern about the way the Diploma has been introduced. It is an innovative and profoundly complex qualification with serious logistical issues to be addressed, yet it the programme of introduction has, according to witnesses, been too fast.⁴²² There is concern that teachers have had little say in how they have been developed. Professor Richard Pring, author of the Nuffield Review of 14–19 education and training, said, “We have got to return to a tradition in which teachers are much more actively involved in creating and thinking about the curriculum rather than—that awful word—‘delivering’ a curriculum created elsewhere.”⁴²³ There is concern about how transportation will work for pupils who have to be moved between different schools. This is especially relevant in rural areas, where long journeys could potentially eat into learning time. We are

418 Q404

419 Ev 162; Q199

420 Ev 72; Ev 105

421 Ev 249

422 Q206

423 <http://education.guardian.co.uk/1419education/story/0,,2277690,00.html>

also concerned about the practicalities of child protection checks on staff in industry who might be working with Diploma pupils. The NAHT has said that this may discourage businesses from wanting to participate.⁴²⁴ What has the Government done to address these concerns?

233. Schools and colleges, who are required to work in collaboration with each other to provide a rounded education for Diploma students, cannot be expected to do so effectively when the accountability regime places them in direct competition with each other. We welcome the introduction of the Diploma and recognise the determination of all concerned to make it work, but we have some concerns about how it will work in a competitive environment.

234. Given its complexity, the Diploma must, in our view, be given an opportunity to settle into its operational phase without undue intervention from the Government. We consider that this is an area best left to the proposed new regulator who we hope will approach Diplomas with a light touch and at a strategic level in the first few years as the initial problems are ironed out over time.

235. The whole education sector would welcome greater clarity on the future direction of Diplomas. We urge the Government to make clear what its intentions are for the future of Diplomas and other 14–19 qualifications and whether it is, in fact, heading towards one, overarching framework for all 14–19 qualifications as Mike Tomlinson’s Working Group on 14–19 Reform proposed in 2004.

The QCA, development and regulation

236. The Government has referred to a perceived conflict of interest inherent in the remit of the QCA. On the one hand, the QCA is responsible for monitoring and advising on the curriculum for children of school age; and for developing associated assessments, tests and examinations. On the other hand, the QCA is the regulator of qualifications offered in schools, colleges and workplaces in England. Others have noted this logical conflict, including the QCA itself, and some have cited weak regulation in some cases; yet there is no serious suggestion that the QCA has acted improperly, especially since the QCA’s development function was hived off to its subsidiary, the National Assessment Agency, in 2004.⁴²⁵ Dr Ken Boston of the QCA told us that:

Our private, but consistent, advice to Government has been that there is a perception that the regulatory decisions could be manipulated by Government, given the way in which we report to Ministers rather than to Parliament.⁴²⁶

[...]the Government have listened to and heard our concerns about the ambiguity present where there is a body that, among other things, is responsible for regulation

424 <http://news.bbc.co.uk/1/hi/education/7379772.stm>

425 Ev 103; Q15; Q55; Q218; written evidence from The Royal Society, section 7

426 Q56

and reports on the maintenance of assessment standards to a Government who are committed to driving up standards to meet particular targets.⁴²⁷

237. The Government's proposal to allocate the development and regulatory functions to two separate agencies is intended to enhance public confidence in standards in the education system.⁴²⁸ The white paper *Confidence in Standards: regulating and developing qualifications and assessment* was published on 17 December 2007 and was jointly presented by the DCSF and the Department for Innovation, Universities and Skills. The white paper proposed two developments:

- a new, independent regulator of tests and qualifications in England, known as the Office of the Qualifications and Examinations Regulator; and
- a new agency to advise Ministers in the monitoring and development of curriculum, assessment and qualifications.

The Office of the Qualifications and Examinations Regulator

238. The independent regulator will report to Parliament through the Children, Schools and Families Committee. It will be the “guardian of standards across the assessment and qualifications system for children, young people and adult learners”, although it will not regulate qualifications awarded by higher education institutions.⁴²⁹

239. The regulator will be responsible for the maintenance of standards over time, which we take to mean ‘assessment standards’ as defined in this Report. As part of this function, the regulator will be required to recognise Awarding Bodies, accredit public qualifications, and monitor and inspect Awarding Bodies.⁴³⁰ It will also regulate National Curriculum tests and moderate assessment at Key Stage 1 and in the Early Years Foundation Stage.⁴³¹

240. Finally, the regulator will oversee the qualifications market and ensure that it is delivering value for money. It will also investigate complaints and consider appeals.⁴³²

Agency for the development of curriculum, assessment and qualifications

241. Under the Government's proposals, the QCA will develop into a new agency, responsible to Ministers, whose main objectives will be:

- to advise Ministers on the monitoring and development of curriculum and related qualifications; on learning and development in early years; and on meeting Government objectives for education and skills; and

427 Q62

428 (2007) DFES & DIUS, *Confidence in Standards: regulating and developing qualifications and assessment*, Cm 7281; Ev 163

429 *ibid.* para 3

430 *ibid.* para 4

431 *ibid.* para 6

432 *ibid.* para 5

- to develop and deliver National Curriculum tests and other forms of assessment; to ensure delivery of public qualifications; and to measure and recognise the achievements of learners and the performance of schools and colleges.

242. It is further proposed that the development agency, rather than the regulator, will develop the criteria for public qualifications, such as GCSE and A-levels, whereas the role of the regulator will be to scrutinise the agency's criteria. The Government intends that the agency will "support the communication of government aims for curriculum and qualifications".⁴³³

243. The work of the QCA to date has been praised by some witnesses.⁴³⁴ The NAHT, for example, said:

The integrity and skill of QCA officials is generally appreciated and respected by the education professionals.⁴³⁵

Others, particularly the Awarding Bodies, have been more critical, stating that regulation has been inconsistent, sometimes overly interventionist and prescriptive.⁴³⁶ Referring to frequent changes to qualifications, Greg Watson of OCR said:

I think that QCA, because of the position it has occupied very close to Government, has tended to find that its role in being a sponsor of change has far outweighed, over time, its responsibility for stability.⁴³⁷

244. Whilst the Government gives the QCA a clear remit for its work, the NAHT states that frustration can arise from the fact that, in its view, the QCA does not have "sufficient freedom in aspects of its work". The NAHT gives the example of the QCA offering "sound professional advice" which the Government has chosen not to follow. At other times, the Government has asked for further investigation to be undertaken when the QCA has recommended caution, for example in relation to the withdrawal of coursework from the GCSE curriculum.⁴³⁸ The NAHT concludes:

QCA is generally effective but there are potential dangers in that it is so strictly controlled by the DfES that all it is empowered to do is offer advice.⁴³⁹

245. Whether the independence of the new regulator will have an impact on the Government's propensity to take advice on regulatory matters remains to be seen. The Government is clear that, in its view, the regulatory functions of the QCA have always been carried out at arm's length from government and the QCA has confirmed this.⁴⁴⁰ Clearly, the new development agency will stand in the shoes of the current QCA in terms of its

433 *ibid.* p6

434 Ev 110; Ev 119; Q13

435 Ev 67

436 Q219; Q220

437 Q219

438 Ev 67

439 Ev 67

440 Q55; Q62

relationship with Government, so that advice on the development side will be given on the same basis as before. There is, therefore, no obvious reason why Government should change its attitude towards advice on development and related matters. However, the new arrangements have broadly been welcomed by witnesses to this inquiry.⁴⁴¹

246. A major rationale for the introduction of an independent regulator is the monitoring and maintenance of assessment standards over time.⁴⁴² Professor Peter Tymms told us that an independent body was essential for this task, a proposition with which Sir Michael Barber agreed.⁴⁴³ Professor Tymms said that standards could not be monitored through the current national testing system due to frequent changes in the curriculum and that an independent body would need to use international standards, as well as the National Curriculum, to track change.⁴⁴⁴ We asked Dr Boston whether there was likely to be anything different about the new regulator which would bring to a halt the drift in assessment standards which he seemed to accept had been a feature of the testing system. Dr Boston replied:

No. The new body—the regulatory authority—will use codes of practice similar to those we have used in the past.⁴⁴⁵

247. OCR have expressed frustration at the annual debate on “standards” which takes place, in their view, at the low level of this year’s papers, the marking of a given paper or the percentage of children awarded a given grade. OCR considers that the debate is taking place at the wrong level and that the focus should really be on the way in which assessment standards are affected by systemic change.

The potential for standards to move and for public confidence to be shaken is greatest when there is wholesale, system-wide change or major structural changes to long-established qualifications. The acid test for looking at the move to an independent regulator is whether we will have a body that is sufficiently able to look at the macro-level changes and the effect that they may have on standards and public confidence and worry much less about the detail of which individual qualification is which.⁴⁴⁶

248. Although there is greater logical consistency in the separation of test development and regulation, this alone is unlikely to address the annual outcry about grade inflation in GCSEs and A-levels. We discussed this with Dr Boston, who thought the new arrangements might help, but admitted that they were unlikely to resolve the issue:

[...] if we consider one of the causes of the August debate to be that the separation of the regulator from Government is not perfectly clear, then that August debate might

441 Ev 103; Ev 146; Q13; Q55; Q173; Q180; Q220

442 (2007) DFES & DIUS, *Confidence in Standards: regulating and developing qualifications and assessment*, Cm 7281; Ev 163

443 Q16

444 Q15

445 Q68

446 Q247

be diminished if the separation were made more apparent. Of course, there may be other issues in the August debate that are not resolved by that situation.⁴⁴⁷

[...] while the basis for [the August debate] might be diminished I am not sure that it is going to go away.⁴⁴⁸

249. We welcome the creation of a development agency and separate, independent regulator on the logical grounds that it is right that development and regulation should be the responsibility of two separate organisations. That assessment standards will now be overseen by a regulator demonstrably free from government control and responsible to Parliament through the Children, Schools and Families Committee is a positive step.

250. However, the Government has failed to address the issue of the standards themselves. In the context of the current testing system, with its ever-changing curriculum and endless test reforms, no regulator, however independent, can assure assessment standards as they are not capable of accurate measurement using the data available. Until the Government allows for standardised sample testing for monitoring purposes, the regulator will be left without the tools required to fulfil its primary function.

447 Q62

448 Q71

6 Conclusion

251. We have been clear that the principle of national testing is sound. However, the central message of our Report has been that national testing can be used in inappropriate ways and that this may lead to damaging consequences for the education system and, most particularly, for children. National testing in England is used for a wide range of purposes, including assessment of pupil attainment, teacher and school accountability and national monitoring. Increasingly, claims are being made that these same tests are also suitable for formative and diagnostic purposes and the new single-level tests are being developed explicitly with this aim. The evidence we have received has been quite clear: a single set of tests cannot validly achieve all of these purposes simultaneously. The purposes of testing must be prioritised and an assessment must be made to establish the extent to which the tests meet the requirements of validity and reliability for each of the identified purposes. This information should then be put in the public domain to give context to the decisions which are made on the basis of published test results and associated statistics.

252. The assumption that the current testing system is capable of meeting validly a wide range of different purposes has distorted the education of some children, which may leave them unprepared for higher education and employment. We consider that the over-emphasis on the importance of national tests, which address only a limited part of the National Curriculum and a limited range of children's skills and knowledge, has resulted in a situation in which many teachers feel compelled to focus unduly on those aspects of the curriculum most likely to be tested and on those students most likely to reach the targets specified by the Government. It is possible to achieve excellent test results by teaching the whole curriculum in a balanced and creative manner, without teaching to the test, but this requires considerable confidence on the part of teachers and schools. In the drive towards more demonstrable reliability in results, teacher assessment and the wider skills of the teaching profession have been undervalued.

253. When the results of national tests are published in the form of performance tables, parents and others are presented with a limited view of a school's activities. We consider that the Government should reform the performance tables to include a wider range of measures of school performance, including results from the most recent Ofsted report, and that this information should be presented in a more accessible manner.

254. As the introduction of the new Diplomas approaches, evidence suggests that teachers feel unprepared for the new qualifications and there is anxiety about the limited amount of training they are due to receive. We wonder how schools will collaborate to provide the new curriculum in the competitive environment created by the imperative to show well in performance tables. Additional problems may arise in relation to the transportation of children between different schools, especially in rural areas; and in relation to the practicalities of child protection checks on businesses working with Diploma students. We look forward to receiving from the Government greater clarity on the future direction of Diplomas.

255. In our view, a brighter future for our education system as a whole lies in a recognition of the professional competence of teachers. The Government should accord a much greater prominence to teacher assessment, which is capable of covering the full curriculum and the

full range of children's knowledge, skills and competences in a way which can never be achieved by a written, externally-marked test. In any reform of the testing system, priority should explicitly be accorded to the purpose of promoting the learning of children. We have been particularly struck by the support in the evidence for the techniques of Assessment for Learning in this respect. Extensive training and ongoing professional support for teachers would be necessary for the success of such a strategy, including the development of a central bank of diagnostic and formative teaching materials which can be administered informally by teachers in classrooms.

256. We emphasise, however, that assessment instruments designed to promote personalised pupil learning, through Assessment for Learning techniques for example, should not be made a part of the accountability regime. This is where we take issue with the single-level tests. The principle of testing when ready may have some merit but, once that system is used for the purposes of school accountability, the focus on effective pupil learning is lost as schools succumb to the imperatives of accountability through targets and performance tables. Looked at from the other direction, tests designed to prioritise the purposes of school accountability and national monitoring cannot simultaneously be suitable for the promotion of personalised pupil learning except at a very shallow level. Such tests cannot possibly attend to the level of detail necessary for planning a pupil's progress through the curriculum on a daily, weekly or monthly basis.

257. We believe that the Government's reforms of the testing system must take account of these concerns if children are to leave school as rounded, knowledgeable, capable individuals ready to progress to further and higher education and contribute effectively to working life.

Appendix

The following table contains a summary of a sample of 10 out of the 20 countries represented in the international comparative analysis of testing and assessment provided to us by the QCA.⁴⁴⁹

Country	Compulsory assessment?	Purpose	Ages of assessment	Subjects assessed	Modes of assessment
Australia	Yes	Evaluative, formative	11/12; 15/16 (sample) 8/9; 10/11; 12/13 (full cohort) 14/15 (full cohort, some states only)	Literacy, numeracy, sciences, civics and citizenship, ICT literacy	Standardised written tests; some practical tasks
Canada	Yes	Evaluative	13 (sample); in future, also 15 (sample)	Reading, mathematics, science	Written tests
England	<i>Yes—Early Years Foundation Stage Profile (from 2008)</i>	<i>Summative</i>	<i>5 (full cohort)</i>	<i>6 statutory ‘early learning goals’</i> <ul style="list-style-type: none"> • <i>Personal, social & emotional development</i> • <i>Communication, language & literacy</i> • <i>Problem solving, reasoning & numeracy</i> • <i>Knowledge & understanding of the world</i> • <i>Physical development</i> • <i>Creative development</i> 	<i>No assessment activities, tasks or tests; teacher observation and assessment recorded against 13 assessment scales derived from ‘early learning goals’</i>
	<i>Yes—National Curriculum assessment</i>	<i>Summative, evaluative</i>	<i>7, 11, 14 (full cohort, publicly funded schools; some independent schools)</i>	<i>Key Stage 1, 2 & 3: English, mathematics Key Stage 2 & 3 only: science</i>	<i>Written tests and tasks; reading tests; mental mathematics tests; statutory teacher assessment</i>

449 Source: QCA memorandum, Annex 5, “Compulsory assessment systems in the INCA countries: thematic probe”, a full version of which is available at <http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilch/memo/169/contents.htm>.

Country	Compulsory assessment?	Purpose	Ages of assessment	Subjects assessed	Modes of assessment
France	Yes—statutory assessment	Diagnostic	8, 11 (full cohort)	Literacy, numeracy	Formal written tests
	Yes—national monitoring assessment	Evaluative	11, 15 (sample)	Various, in a cycle starting with reading (2003); foreign languages (2004); history/geography and civics (2006); science and technology (2007); mathematics (2008); returning to reading.	Standardised written tests
Germany	Under development	Evaluative	10, 15/16 (sample)	Age 10: German, mathematics Age 15/16: German, mathematics, first foreign language (English or French)	Information not yet available
Japan	Periodic Nationwide Academic Ability Assessment	Evaluative, informative	11/12, 14/15 (full cohort)	Japanese, mathematics, 'eagerness to learn', 'daily life habits'	Written tests
	Yes—end of secondary achievement tests	Summative, informative	15 (full cohort)	Japanese, social studies, mathematics, science and English (depending on the administering prefecture)	Formal written tests (usually)
New Zealand	Optional school entry assessment	Diagnostic, formative, informative	5/6	Literacy, numeracy, oral language	Observation of tasks
	Yes—National Education Monitoring Project	Evaluative	8/9, 12/13 (sample)	All curriculum areas over a 4-year period; also knowledge, skills, motivation and attitudes.	Assessment tasks
Spain	Yes—national sample surveys (current)	Evaluative, informative	12, 16 (sample)	Mathematics, Spanish (language & literature) and regional languages, natural and social sciences, geography and history, foreign languages.	Formal written tests (mostly)
	Diagnostic evaluations (proposed)	Diagnostic, formative	10/11, 13/14 (full cohort)	Key competences	Written tests (likely)

Country	Compulsory assessment?	Purpose	Ages of assessment	Subjects assessed	Modes of assessment
Sweden	Yes	Diagnostic, evaluative	9, 12, 14 (voluntary) 16 (compulsory)	9: literacy (Swedish), numeracy 12, 14, 16: Swedish/Swedish as a second language, English, mathematics	Written and oral tests
USA	National Assessment of Educational Progress	Evaluative, informative	9/10, 13/14, 17/18 (sample)	Reading, mathematics, science, writing, American history, civics, geography, arts	Written tests (constructed-response and multiple-choice questions)
	'No Child Left Behind' assessments	Evaluative	8-14 (full cohort)	Literacy, numeracy, science	Written tests (usually)
	A variety of state-specific tests				

Conclusions and recommendations

The need for national testing

1. We consider that the weight of evidence in favour of the need for a system of national testing is persuasive and we are content that the principle of national testing is sound. Appropriate testing can help to ensure that teachers focus on achievement and often that has meant excellent teaching, which is very welcome. (Paragraph 25)

The purposes of national testing

2. The evidence we have received strongly favours the view that national tests do not serve all of the purposes for which they are, in fact used. The fact that the results of these tests are used for so many purposes, with high-stakes attached to the outcomes, creates tensions in the system leading to undesirable consequences, including distortion of the education experience of many children. In addition, the data derived from the testing system do not necessarily provide an accurate or complete picture of the performance of schools and teachers, yet they are relied upon by the Government, the QCA and Ofsted to make important decisions affecting the education system in general and individual schools, teachers and pupils in particular. In short, we consider that the current national testing system is being applied to serve too many purposes. (Paragraph 44)
3. We consider that the over-emphasis on the importance of national tests, which address only a limited part of the National Curriculum and a limited range of children's skills and knowledge has resulted in teachers narrowing their focus. Teachers who feel compelled to focus on that part of the curriculum which is likely to be tested may feel less able to use the full range of their creative abilities in the classroom and find it more difficult to explore the curriculum in an interesting and motivational way. We are concerned that the professional abilities of teachers are, therefore, under-used and that some children may suffer as a result of a limited educational diet focussed on testing. We feel that teacher assessment should form a significant part of a national assessment regime. As the Chartered Institute of Educational Assessors states, "A system of external testing alone is not ideal and government's recent policy initiatives in progress checks and diplomas have made some move towards addressing an imbalance between external testing and internal judgements made by those closest to the students, i.e. the teachers, in line with other European countries". (Paragraph 58)
4. We are concerned about the Government's stance on the merits of the current testing system. We remain unconvinced by the Government's assumption that one set of national tests can serve a range of purposes at the national, local, institutional and individual levels. We recommend that the Government sets out clearly the purposes of national testing in order of priority and, for each purpose, gives an accurate assessment of the fitness of the relevant test instrument for that purpose, taking into account the issues of validity and reliability. (Paragraph 61)

5. We recommend further that estimates of statistical measurement error be published alongside test data and statistics derived from those data to allow users of that information to interpret it in a more informed manner. We urge the Government to consider further the evidence of Dr Ken Boston, that multiple test instruments, each serving fewer purposes, would be a more valid approach to national testing. (Paragraph 62)

Performance targets and tables

6. We endorse the Government's view that much can and should be done to assist children who struggle to meet expected standards. However, we are concerned that the Government's target-based system may actually be contributing to the problems of some children. (Paragraph 81)
7. We believe that the system is now out of balance in the sense that the drive to meet government-set targets has too often become the goal rather than the means to the end of providing the best possible education for all children. This is demonstrated in phenomena such as teaching to the test, narrowing the curriculum and focussing disproportionate resources on borderline pupils. We urge the Government to reconsider its approach in order to create incentives to schools to teach the whole curriculum and acknowledge children's achievements in the full range of the curriculum. The priority should be a system which gives teachers, parents and children accurate information about children's progress. (Paragraph 82)
8. Whilst we consider that Contextualised Value Added scores are potentially a valuable addition to the range of information available to parents and the public at large when making judgments about particular schools, we recommend that the information be presented in a more accessible form, for example graphically, so that it can more easily be interpreted. (Paragraph 98)
9. We are concerned about the underlying assumptions on which Contextualised Value Added scores are based. Whilst it may be true that the sub-groups adjusted for in the Contextualised Value Added measure may statistically perform less well than other sub-groups, we do not consider that it should be accepted that they will always perform less well than others. (Paragraph 99)
10. In addition to these specific recommendations about Contextual Value Added scores, we recommend that the Government rethinks the way it publishes the information presented in the Achievement and Attainment Tables generally. We believe that this information should be presented in a more accessible manner so that parents and others can make a holistic evaluation of a school more easily. In addition, there should be a statement with the Achievement and Attainment Tables that they should not be read in isolation, but in conjunction with the relevant Ofsted report in order to get a more rounded view of a school's performance and a link to the Ofsted site should be provided. (Paragraph 100)
11. The scope of this inquiry does not extend to a thorough examination of the way Ofsted uses data from the performance tables under the new, lighter touch, inspection regime. However, we would be concerned if Ofsted were, in fact, using

test result data as primary inspection evidence in a disproportionate manner because of our view that national test data are evidence only of a very limited amount of the important and wide-ranging work that schools do. (Paragraph 102)

12. We consider that schools are being held accountable for only a very narrow part of their essential activities and we recommend that the Government reforms the performance tables to include a wider range of measures, including those from the recent Ofsted report. (Paragraph 106)

The consequences of high-stakes uses of testing

13. We received substantial evidence that teaching to the test, to an extent which narrows the curriculum and puts sustained learning at risk, is widespread. Whilst the Government has allocated resources to tackle this phenomenon and improve practice they fail to accept the extent to which teaching to the test exists and the damage it can do to a child's learning. We have no doubt that teachers generally have the very best intentions in terms of providing the best education they can for their pupils. However, the way that many teachers have responded to the Government's approach to accountability has meant that test results are pursued at the expense of a rounded education for children. (Paragraph 130)
14. We believe that teaching to the test and this inappropriate focus on test results may leave young people unprepared for higher education and employment. We recommend that the Government reconsiders the evidence on teaching to the test and that it commissions systematic and wide-ranging research to discover the nature and full extent of the problem. (Paragraph 131)
15. A creative, linked curriculum which addresses the interests, needs and talents of all pupils is the casualty of the narrow focus of teaching which we have identified. Narrowing of the curriculum is problematic in two ways: core subjects are emphasised to the detriment of other, important elements of the broader curriculum; and, for those subjects which are tested in public examinations, the scope and creativity of what is taught is compromised by a focus on the requirements of the test. We are concerned that any efforts the Government makes to introduce more breadth into the school curriculum are likely to be undermined by the enduring imperative for schools, created by the accountability measures, to ensure that their pupils perform well in national tests. (Paragraph 140)
16. We acknowledge the reforms the Government has made to GCSE and A-level examinations. However, the Government must address the concerns expressed by witnesses, among them Dr Ken Boston of the QCA, who see the burden of assessment more in terms of the amount of time and effort spent in preparation for high-stakes tests than in the time taken to sit the tests themselves. This could be achieved by discouraging some of the most inappropriate forms of preparation and reducing the number of occasions on which a child is tested. (Paragraph 149)
17. We are persuaded by the evidence that it is entirely possible to improve test scores through mechanisms such as teaching to the test, narrowing the curriculum and concentrating effort and resources on borderline students. It follows that this

apparent improvement may not always be evidence of an underlying enhancement of learning and understanding in pupils. (Paragraph 161)

18. We consider that the measurement of standards across the full curriculum is virtually impossible under the current testing regime because national tests measure only a small sample of pupils' achievements; and because teaching to the test means that pupils may not retain, or may not even possess in the first place, the skills which are supposedly evidenced by their test results. (Paragraph 162)
19. It is not possible for us to come to a definitive view on grade inflation in the context of such a wide-ranging inquiry. However, it seems clear to us from the evidence that we have received that the Government has not engaged with the complexity of the technical arguments about grade inflation and standards over time. We recommend that the Government addresses these issues head-on, starting with a mandate to the QCA or the proposed new regulator to undertake a full review of assessment standards. (Paragraph 171)
20. Whilst we do not doubt the Government's intentions when it states that "The National Curriculum sets out a clear, full and statutory entitlement to learning for all pupils, irrespective of background or ability", we are persuaded that in practice many children have not received their entitlement and many witnesses believe that this is due to the demands of national testing. (Paragraph 183)
21. We are persuaded that the current system of national tests should be reformed in order to decouple the multiple purposes of measuring pupil attainment, school and teacher accountability and national monitoring. The negative impacts of national testing arise more from the targets that schools are expected to achieve and schools' responses to them than from the tests themselves. (Paragraph 184)
22. School accountability should be separated from this system of pupil testing, and we recommend that the Government consult widely on methods of assuring school accountability which do not impact on the right of children to a balanced education. (Paragraph 185)
23. We recommend that the purpose of national monitoring of the education system, particularly for policy formation, is best served by sample testing to measure standards over time and that cohort testing is neither appropriate nor, in our view, desirable for this purpose. We recommend further that, in the interests of public confidence, such sample testing should be carried out by a body at arms length from the Government and suggest that it is a task either for the new regulator or a body answerable to it. (Paragraph 186)

Single-level tests

24. Our predecessors warned the Government about bringing in new tests with undue haste. We recommend that the Government allows sufficient time for a full pilot of the new single-level tests and ensures that any issues and problems arising out of that pilot are fully addressed before any formal roll-out of the new regime to schools. (Paragraph 198)

25. *Making Good Progress* characterises single-level tests as integral to personalised learning and Assessment for Learning yet also the means by which to generate summative data. We agree with the National Foundation for Educational Research that this single assessment instrument cannot validly perform these functions simultaneously and, if it is attempted, there is a danger that the single-level tests will work for neither purpose. The single-level tests may be useful, however, if their purpose is carefully defined and the tests are developed to ensure they are valid and reliable specifically for those purposes. (Paragraph 210)
26. We recommend that, if single-level tests are introduced, they are used for summative purposes only and that Assessment for Learning and personalised learning are supported separately by enhanced professional development for teachers, backed up with a centralised bank of formative and diagnostic assessment materials on which teachers can draw as necessary on a regular basis. (Paragraph 211)
27. Single-level tests may have some positive effects and we certainly approve of the Government's new emphasis on the personalised approach. However, the Government has structured the single-level testing system in such a way as to risk a transposition of existing, systemic problems into the new arrangements. Without structural modification, we foresee that the existing problems—including teaching to the test, narrowing of the taught curriculum and the focus on borderline candidates to the detriment of others—will continue under the single-level test regime. (Paragraph 215)
28. We believe that true personalised learning is incompatible with a high-stakes single-level test which focuses on academic learning and does not assess a range of other skills which children might possess. Children who struggle with the core subjects may receive more targeted assistance in those subjects. However, if this means that children who are struggling with core subjects get less opportunity to access the wider curriculum, they risk being put off learning at an early age. We call upon the Government to invest in ways to help and, if necessary, train teachers to improve the basic skills of struggling pupils while enhancing their enjoyment of learning and guaranteeing their access to a broad curriculum. (Paragraph 216)
29. We are concerned about the “one-way ratchet” on the attainment of test levels under the single-level testing regime and we find persuasive the evidence that this may lead to an apparent, but artificial, improvement in performance standards. We recommend that the Government consider further whether it is in children's best interests that they may be certified to have achieved a level of knowledge and understanding which they do not, in truth, possess. We suspect that this may lead to further disillusionment and children perceiving themselves as ‘failures’. (Paragraph 217)
30. We recommend that the Government urgently rethinks its decision to use progression targets, based on pupils' achievement in single-level tests, for the purposes of school accountability. If such high-stakes accountability measures are combined with more frequent testing of children, the negative effect on children's education experiences promises to be greater than it is at present. We urge the Government to listen to the QCA, which has already warned of the dangers of

saddling the single-level tests with the same range of purposes which the Key Stage tests demonstrably cannot bear. (Paragraph 218)

Diplomas

31. We welcome the Government's stated intentions that both the vocational and the general elements of Diplomas should be reflected in the methods of assessment used. We caution the Government against any haste in shifting this delicate balance in future until the full implications of such a shift have been understood. (Paragraph 225)
32. Schools and colleges, who are required to work in collaboration with each other to provide a rounded education for Diploma students, cannot be expected to do so effectively when the accountability regime places them in direct competition with each other. We welcome the introduction of the Diploma and recognise the determination of all concerned to make it work, but we have some concerns about how it will work in a competitive environment. (Paragraph 233)
33. Given its complexity, the Diploma must, in our view, be given an opportunity to settle into its operational phase without undue intervention from the Government. We consider that this is an area best left to the proposed new regulator who we hope will approach Diplomas with a light touch and at a strategic level in the first few years as the initial problems are ironed out over time. (Paragraph 234)
34. The whole education sector would welcome greater clarity on the future direction of Diplomas. We urge the Government to make clear what its intentions are for the future of Diplomas and other 14–19 qualifications and whether it is, in fact, heading towards one, overarching framework for all 14–19 qualifications as Mike Tomlinson's Working Group on 14–19 Reform proposed in 2004. (Paragraph 235)

Regulation and development: the new arrangements

35. We welcome the creation of a development agency and separate, independent regulator on the logical grounds that it is right that development and regulation should be the responsibility of two separate organisations. That assessment standards will now be overseen by a regulator demonstrably free from government control and responsible to Parliament through the Children, Schools and Families Committee is a positive step. (Paragraph 249)
36. However, the Government has failed to address the issue of the standards themselves. In the context of the current testing system, with its ever-changing curriculum and endless test reforms, no regulator, however independent, can assure assessment standards as they are not capable of accurate measurement using the data available. Until the Government allows for standardised sample testing for monitoring purposes, the regulator will be left without the tools required to fulfil its primary function. (Paragraph 250)

Formal Minutes

Wednesday 7th May 2008

Morning sitting

Members present:

Mr Barry Sheerman, in the Chair

Annette Brooke
Mr Douglas Carswell
Mr David Chaytor
Paul Holmes

Fiona Mactaggart
Andy Slaughter
Mr Graham Stuart

Testing and Assessment

The Committee considered this matter.

Draft Report, proposed by the Chairman, brought up and read.

Draft Report, proposed by Mr Douglas Carswell, brought up and read, as follows:

“No need for State-run testing

1. In the late 1980s there was a *de facto* nationalisation of testing with the introduction of national criteria for GCSEs and, more generally, the introduction of the National Curriculum. From the evidence available to us it is clear that this system has been a failure.

2. The results of GCSEs and national curriculum tests have become debased currency, because of grade inflation and consequent lack of validity (that is, the judgments that someone might make about the capabilities of a pupil reaching a certain level are undermined by, amongst other things, variations in pass marks over time and teaching to the test). By moving away from ‘hard’ results by using contextual value added (CVA) scores in national curriculum tests certainty is lost along with reliability (that is, confidence that pupils reaching the same level of performance gain the same outcome). Furthermore the outcomes, particularly the CVA scores, are incomprehensible to parents, employers and other end-users.

3. The consequence of the failure is stark:

“Sixteen per cent of the adult workforce is illiterate. Five million adults have no qualifications. A million teenagers have failed to gain even the lowest grade in five GCSEs since Labour came to power, and 23,000 pupils earned no GCSEs at all.”⁴⁵⁰

These figures show the extent of the crisis in education, which centralised testing has failed to arrest or reverse. Indeed, given the extent of grade inflation and the relativism of outcomes from CVA measures, the testing system excuses and legitimises failure. CVA enshrines in education policy the odious assumption that socio-economic background at birth directly determines life outcomes.

4. Much of the evidence we received argued that we need the state run system. David Bell, Permanent Secretary at the Department of Children, Schools and Families, told us:

⁴⁵⁰ *Time to crush the NUT*, George Bridges, *Daily Telegraph*, 22 April 2008.

“I do not accept that we can ever have a system without good and robust national testing and public examinations, the results of which are made available to the public.”⁴⁵¹

Similarly, the former DfES in its written submission to the inquiry said:

“The benefits brought about by [National Curriculum testing], compared to the time before the accountability of the National Curriculum, have been immense. The aspirations and expectations of pupils and their teachers have been raised. For parents, the benefits have been much better information not only about the progress their own child is making but also about the performance of the school their child attends. And for the education system as a whole, standards of achievement have been put in the spotlight, teachers’ efforts have been directed to make a difference and performance has improved. The public has a right to demand such transparency at a time of record investment in education.”⁴⁵²

5. This technocratic insistence that state run testing is required to improve education in schools is the same argument used to justify state involvement in airlines and telecommunications in the 1970s. In fact the system produces the reverse effect to that claimed for it. It fails to provide proper accountability as it is too complex and can lead people to make poor choices, and it fails to provide quality assurance because the outcomes, being contextualised and relativised, lack rigour. It was, however, no surprise that many of the witnesses argued that there is a continuing need for state run testing as so many of them earn their living from that system.

6. We consider that the Qualifications and Curriculum Authority (QCA) is a significant part of the problem. It is an unaccountable quango that has presided over the massive grade inflation that has led to loss of confidence in what the results of exams of all kinds actually signify. Given our belief that the state should not run the testing system, we consider the role of the QCA should be abolished. If that does not happen, the House of Commons should at least be given the power to ratify senior appointments and approve its budget. Senior staff should also be on fixed term contracts.

7. For the same reasons we reject the Government’s plan to establish two new quangos, a development agency and a separate regulator. The establishment of more so-called arms length bodies has been the default option for politicians for the past generation, but we are in no doubt that these bodies have been very much part of the problem in education and cannot be part of the solution.

Plurality of tests

8. Dr Ken Boston of the QCA argued that it would be wrong to have a large number of different tests:

“My judgment is that, given that there are so many legitimate purposes of testing, and [a paper prepared by the QCA] lists 22, it would be absurd to have 22 different sorts of tests in our schools. However, one serving 14 purposes is stretching it too far. Three or four serving three or four purposes each might get the tests closer to what they were designed to do. ... when you put all of these functions on one test, there is the risk that you do not perform any of those functions as perfectly as you might. What we need to do is not to batten on a whole lot of functions to a test, but restrict it to three or four prime functions that we believe are capable of delivering well.”

⁴⁵³

451 Q327

452 Ev 157

453 Q7

We disagree strongly with this view; there is a real need for plurality of provision for testing. This would allow for **innovation**, testing things which are not yet envisaged, and **choice**, of different standards and degrees of rigour. This will not be achieved with a state run system testing things the state wishes to test.

Purpose of testing

9. **The purpose of testing ought to be identify levels of attainment before a pupil moves on to the next stage of education or into employment.** This testing is best left to those civic institutions which act as gatekeepers for that 'next stage'; universities, professions, schools and employers. Some of the current purposes to which tests are put inevitably distort outcomes. These includes testing as a means of gauging social engineering by measuring the impact of particular Government policies, and requiring schools to achieve targets of a certain level of achievement amongst their pupils, which leads to the unintended consequences of teaching to the test, narrowed curriculum and other distortions.

10. There is no doubt that we do need standard tests available across the country, but these should not be state-run. There are good examples to be found amongst the professions, such as law, medicine and veterinary medicine.

11. **This report is not anti-testing, as there is an undoubted need for testing that identifies levels of educational attainment. We therefore have no sympathy with those teacher unions, for example, which oppose all testing, as we consider that to be merely seeking to avoid accountability. The decisions about how, when and what to test, however, should be left to autonomous schools and universities. This would require schools and universities to have legally enshrined autonomy over admissions, complemented by a legally enshrined right of parents to choose the school their children attend. Removing the influence of the state from these processes and decisions is the only way to achieve improved educational attainment in the future.** “

Motion made, and Question proposed, That the Chairman's draft report be read a second time, paragraph by paragraph.—(*The Chairman.*)

Amendment proposed, to leave out the words “Chairman's draft report” and insert the words “draft report proposed by Mr Douglas Carswell”.—(*Mr Douglas Carswell.*)

Question put, that the Amendment be made.

The Committee divided.

Ayes, 1

Mr Douglas Carswell

Noes, 4

Annette Brooke
Mr David Chaytor
Fiona Mactaggart
Mr Graham Stuart

Ordered, That the Chairman's draft Report be read a second time, paragraph by paragraph.

Ordered, That further consideration of the Chairman's draft report be now adjourned.

Report to be further considered this day.

[Adjourned till this day at 2.30 pm

Wednesday 7th May 2008

Afternoon sitting

Members present:

Mr Barry Sheerman, in the Chair

Annette Brooke
Mr David Chaytor
John Heppell
Sharon Hodgson
Paul Holmes

Fiona Mactaggart
Andy Slaughter
Mr Graham Stuart
Lynda Waltho

Testing and Assessment

Consideration of Chairman's draft report resumed.

Paragraphs 1 to 257 read and agreed to.

Summary agreed to.

Resolved, That the Report be the Third Report of the Committee to the House.

Ordered, That the Chairman do make the Report to the House.

Ordered, That embargoed copies of the report be made available, in accordance with the provisions of Standing Order No. 134.

Written evidence was ordered to be reported to the House for printing with the report, together with written evidence reported and ordered to be published on 4 July 2007 and 12 March 2008.

Written evidence was ordered to be reported to the House for placing in the Library and Parliamentary Archives.

[Adjourned till Monday 12 May at 3.30 pm

Witnesses (Volume II)

Monday 10 December 2007

Page

Professor Sir Michael Barber, Expert Partner, Global Public Sector Practice, McKinsey and Company, and **Professor Peter Tymms**, Director of Curriculum, Evaluation and Management, School of Education, Durham University Ev 1

Monday 17 December 2007

Dr Ken Boston, Chief Executive, Qualifications and Curriculum Authority (QCA) Ev 34

Monday 14 January 2008

Dr Mary Bousted, General Secretary, Association of Teachers and Lecturers (ATL), **Mick Brookes**, General Secretary, National Association of Head Teachers (NAHT), **Brian Lightman**, President, Association of School and College Leaders (ASCL), and **Keith Bartley**, Chief Executive, General Teaching Council for England (GTC) Ev 86

Monday 21 January 2008

Dr Andrew Bird, Deputy Director General, Assessment and Qualifications Alliance (AQA), **Murray Butcher**, Director, Assessment and Quality, City & Guilds, **Jerry Jarvis**, Managing Director, Edexcel, and **Greg Watson**, Chief Executive, Oxford, Cambridge and RSA Examinations (OCR) Ev 123

Monday 28 January 2008

Professor Steve Smith, Vice-Chancellor, University of Exeter, **Professor Madeleine Atkins**, Vice-Chancellor, University of Coventry, and **Susan Anderson**, Director, Human Resources Policy, and **Richard Wainer**, Principal Policy Adviser, Education and Skills, Confederation of British Industry (CBI) Ev 146

David Bell, Permanent Secretary, **Sue Hackman**, Chief Adviser on School Standards, and **Jon Coles**, Director, 14–19 Reform, Department for Children, Schools and Families (DCSF) Ev 166

Monday 18 February 2008

Jim Knight MP, Minister for Schools and Learners, and **Ralph Tabberer**, Director General, Schools Directorate, Department for Children, Schools and Families Ev 177

List of written evidence printed (Volume II)

1	Dr Ken Boston, Chief Executive, Qualifications and Curriculum Authority (QCA)	Ev 21
2	Association of School and College Leaders (ASCL)	Ev 46: Ev 55
3	Association of Teachers and Lecturers (ATL)	Ev 55
4	National Association of Head Teachers (NAHT)	Ev 66
5	General Teaching Council for England (GTC)	Ev 73
6	Assessment and Qualifications Alliance (AQA)	Ev 101
7	City & Guilds	Ev 109
8	Edexcel	Ev 113
9	OCR	Ev 118
10	Confederation of British Industry (CBI)	Ev 143
11	Department for Education and Skills (DfES)	Ev 157
12	Department for Children, Schools and Families (DCSF)	Ev 163
13	Jim Knight MP, Minister of State for Schools and Learners, Department for Children, Schools and Families (DCSF)	Ev 176: Ev 196
14	Association of Colleges (AoC)	Ev 198
15	Paul Black, Emeritus Professor of Education King's College London, John Gardner, Professor of Education Queen's University Belfast, and Dylan Wiliam, Professor and Deputy Director, Institute of Education, London	Ev 202
16	Cambridge Assessment	Ev 205
17	Chartered Institute of Educational Assessors (CIEA)	Ev 222
18	Institute for Public Policy Research (IPPR)	Ev 232
19	National Association of Schoolmasters Union of Women Teachers (NASUWT)	Ev 244
20	National Foundation for Educational Research (NFER)	Ev 250
21	National Union of Teachers (NUT)	Ev 261
22	Professor Colin Richards	Ev 267
23	The Wellcome Trust	Ev 269
24	Hampshire County Council	Ev 270

List of unprinted written evidence

Written evidence has been published on our website at:

<http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilisch/memo/169/contents.htm>

Written evidence has also been reported to the House. Copies have been placed in the House of Commons Library, where they may be inspected by Members. Other copies are in the Parliamentary Archives, and are available to the public for inspection. Requests for inspection should be addressed to The Parliamentary Archives, Houses of Parliament, London SW1A 0PW (tel. 020 7219 3074). Opening hours are from 9.30 am to 5.00 pm on Mondays to Fridays.

TA02	Heading for Inclusion, Alliance for Inclusive Education
TA05	Portway Infant School, Hampshire
TA06	Purbrook Junior School, Waterlooville
TA07	Epilepsy Action

TA08	Doug French, University of Hull
TA10	LexiaUK
TA14	Advisory Committee on Mathematics Education (ACME)
TA15	Mathematics in Engineering and Industry (MEI)
TA17	RNIB
TA19	Institute of Physics
TA20	Association for Achievement and Improvement through Assessment (AAIA)
TA21	Campaign for Science and Education (CaSE)
TA22	Dr A Gardiner, University of Birmingham
TA25	Barbara J Cook, Headteacher, Guillemont Junior School, Farnborough, Hampshire
TA26	Alliance for Inclusive Education
TA27	The Royal Society
TA28A	Wellcome Trust, Annex B
TA30	Ms Janet Graham, Director, Supporting Professionalism in Admissions Programme (SPA)
TA31	Qualification and Curriculum Authority, Annexes 3-5
TA33	Lorraine Smith, Headteacher, Western Church of England Primary School, Winchester, Hampshire
TA34	Association of Science Education (ASE)
TA35	The Mathematical Association
TA37	Richard Cooper
TA38	S Forrest, Teacher of Mathematics, Wokingham Authority
TA39	H Nickels, Headteacher, Silverton CE Primary School, Devon
TA42	SCORE—Science Community Partnership Supporting Education
TA45	University of Cambridge International Examinations
TA48	Warwick Mansell

List of Reports from the Committee during the current Parliament

The reference number of the Government's response to each Report is printed in brackets after the HC printing number.

Session 2007–08

First Report	Special	Creative Partnerships and the Curriculum: Government Response to the Eleventh Report from the Education and Skills Committee, Session 2006–07	HC 266
Second Report	Special	Special Educational Needs: Assessment and Funding: Government Response to the Tenth Report from the Education and Skills Committee, Session 2006–07	HC 298
First Report		Children and Young Persons Bill [Lords]	HC 359
Second Report		The Department for Children, Schools and Families and the Children's Plan	HC 213