Independent Review of Key Stage 2 testing, assessment and accountability

Final Report

Lord Bew

June 2011

Contents

Remit of the Review into Key Stage 2 testing, assessment and accountability

This Review has been framed by a particular remit. It may be helpful to begin by revisiting our remit so that we can be clear about what is in our scope.

There are two broad positions to which the Secretary of State has asked us to adhere throughout. Firstly, the Government is mindful that the Organisation for Economic Cooperation and Development (OECD) concludes that external accountability is a key driver of improvement in education and particularly important for the least advantaged. It therefore views a system of objectively measuring pupil progress and holding schools to account as vital. Secondly, the Government has made it clear that it wants schools and teachers to be free to set their own direction, trusted to exercise their professional discretion and accountable for the progress of the children in their care. The Secretary of State has therefore been clear that school autonomy must be accompanied by robust accountability.

Within those parameters, this Review was asked to address the following key issues:

- how best to ensure that the system of assessment in primary schools can improve standards of attainment and progress of pupils, and help narrow gaps;

- how best to ensure that schools are properly and fairly accountable to pupils, parents and the taxpayer for the achievement and progress of every child, on the basis of objective and accurate assessments; and that this reflects the true performance of the school;

- how to avoid, as far as possible, the risk of perverse incentives, over-rehearsal and reduced focus on productive learning;

- how to ensure that parents have good quality information on the progress of their children and the success of schools;

- how to ensure that performance information is used and interpreted appropriately within the accountability system by other agencies, increasing transparency and preserving accountability to parents, pupils and the taxpayer, while avoiding the risk of crude and narrow judgements being made;

- how to ensure that tests are rigorous, and as valid and reliable as possible, within an overall system of assessment (including teacher assessment) which provides the best possible picture of every child's progress;

- how best to ensure that the assessment system allows us to make comparisons with education systems internationally;

- how to make administration of the system as simple and cost-effective as possible, with minimal bureaucracy.

Membership of the Review Panel

Lord Bew is a cross-bench peer, Professor of Irish Politics at Queen's University in Belfast, and a Member of the Royal Irish Academy. He was a historical adviser to the Saville Inquiry from 1998 to 2001.

Membership of the panel in full is:

- Lord Bew – Chair.
- Helen Clegg OBE – Executive Head teacher, Shiremoor Primary School in North Tyneside. National Leader of Education.
- Sally Coates – Principal, Burlington Danes Academy in West London.
- Kate Dethridge – Head teacher, Churchend Primary School in Reading. National Leader of Education.
- Lubna Khan – Head teacher, Berrymede Junior School in Ealing. Local Leader of Education.
- Ruth Miskin – Founder, Read-Write Inc. and former primary head teacher.
- Miriam Rosen – Former Executive Director, Ofsted.
- Tim Sherriff – Head teacher, Westfield Community School in Wigan. Local Leader of Education.
- Greg Wallace – Executive Principal of Best Start Federation in Hackney. National Leader of Education.

Representatives of Ofsted and Ofqual act as observers.

Outline of the evidence-gathering process

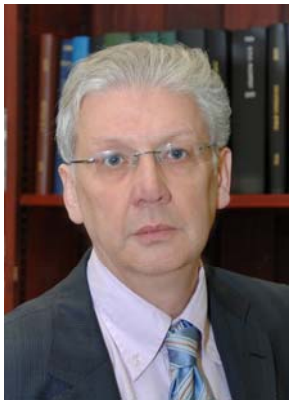
Given the scale and complexity of this Review, we endeavoured to gather as much evidence and feedback as possible in an open, transparent and outward-facing way. A 12-week call for evidence invited all interested parties to contribute. As a result we have received around 4,000 online responses, taken oral evidence from 50 stakeholders, and many written submissions have been sent in. A list of stakeholders is included at the end of this Report.

In early April we published our *Progress Report*, which outlined the range of evidence and opinion received by the Review. In particular, it outlined the main views we heard through oral evidence sessions and written submissions. It did not attempt to draw conclusions or make recommendations, which is the task of this Final Report.

We have also published a *Call for Evidence Report*, which summarises the main findings from the nearly 4,000 respondents to the Review's call for evidence.

This *Final Report* discusses the evidence, including the published research material we have considered, outlines the conclusions we have reached and sets out our recommendations.

Foreword by Lord Bew



The system of testing, assessment and accountability at the end of Key Stage 2 has a profound impact on primary school education. It is crucial for pupils, parents, teachers, head teachers and the nation as a whole that the system works effectively and is fair for all involved.

Given the scale of the challenge and the importance of getting Key Stage 2 testing, assessment and accountability right, the Secretary of State asked us seven months ago to conduct an independent review of the system and to set out our recommendations for how it can be improved.

We are grateful to everyone who has contributed to the Review. We very much appreciate the 4,000 responses to the online call for evidence, the evidence from 50 stakeholders who have spoken to us in formal sessions, and the many written submissions we have received. We are fortunate that this Review has generated an immense amount of evidence and feedback in this area, which we have been careful to take into account fully.

It has been a real privilege to hear from so many colleagues who are so passionately committed to improving the current system. We are grateful for the open and constructive way in which so many people have explained their views and presented evidence and feedback.

There is clearly much about primary school education in this country that we should celebrate and of which we should be proud. To take just one example, the attainment gap between pupils known to be eligible for free school meals and their peers, while still too large, has fallen by 7 percentage points in English and 6 percentage points in mathematics between 2002 and 2010.

We have been particularly impressed by the outstanding commitment of teachers and head teachers across the country about whom we have heard during the Review. We understand how hard they are working, often in very challenging contexts, to give children in their care the best possible education.

However, in relation to the area of primary school education which falls within our remit – Key Stage 2 testing, assessment and accountability – significant concerns have frequently been raised. Many of those who gave evidence have highlighted positive aspects of the current system and said that they should be retained; many other respondents have argued that parts of the system are unhelpful, and at worst a barrier to teaching and learning. The strength of feeling associated with these concerns shows that some change is clearly needed. A commitment to standards should not, in principle, be at the expense of creativity.

The range of opinion in this area has been striking. It is inevitable, given the strength of feeling and the wide range of views, that much of the evidence and feedback is conflicting or contradictory. However, we have been surprised that every suggestion has generated substantial drawbacks and risks, and that every proposal which enjoys any significant support from some respondents can prompt a negative reaction from others. As many respondents have agreed, there is no single, simple solution to this difficult problem.

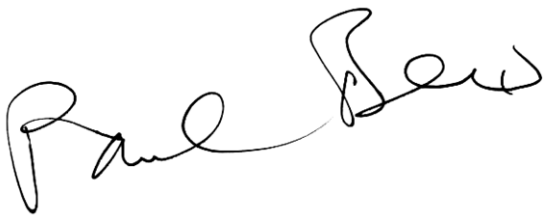
It became clear early on that, based on the range of evidence and opinion and the complexity of the issues in our remit, it would not be possible for this Review to recommend a series of solutions which would command universal support.

However, this has presented a real opportunity to make recommendations based purely on what is educationally the right approach. This is not a subject where finding the middle ground between differing opinions would be enough. Simply reaching a compromise between the different views would not do justice to pupils, parents, teachers, and head teachers, who will want to be absolutely confident that this part of the system is right.

Our aim throughout has been to develop a system which best supports the learning of every pupil in this country. Individual schools, teachers, head teachers and parents clearly play defining roles in ensuring each child learns and achieves his or her potential. We are aware that it is good teaching which raises standards and good teachers seek to know each child as an individual and work in partnership with him or her to support his or her learning. We believe that the Government should free teachers up as much as possible to do so. However, we also believe that the right national system can play an essential part in helping to support the learning of every child, and particularly in making sure that no one is left behind.

It may be worth noting at this stage that, while improving an individual school or class can be done through specific and tailored changes, system-wide improvement will always require a broader set of changes which must work for all schools. We have made system-wide recommendations, which apply to around 17,000 primary schools of enormously varying size, intake and context. It is therefore possible that individual pupils, parents, teachers or head teachers may not appreciate immediately how these system-wide recommendations apply to them or their school. This kind of disjoint happens to an extent in all large systems.

However, we would like to be quite clear that throughout this process we have always focused on how best to support the learning of each individual child. We believe that our recommendations, when put together as a package, will improve the system to deliver fairer and more effective Key Stage 2 testing, assessment and accountability for all pupils, parents, teachers and head teachers.



Lord Paul Bew

Executive Summary

There are certain principles that underpin this Final Report, which it may be helpful to summarise before we move on to our final recommendations.

Key principles which underpin our recommendations

The evidence submitted to the Review overwhelmingly accepts that, as public bodies, **schools should be held accountable for the education of their pupils**. This is an important point of principle which we wholeheartedly support.

There is widespread research evidence which suggests that **external school-level accountability is important in driving up attainment and pupils' progress**, and we find this evidence compelling. We understand that, at school and pupil level, this can lead to what can seem like frustrating pressure and an unnecessarily 'high stakes' system, but we believe that holding each school accountable externally is essential.

However, **external school-level accountability is only acceptable if it is fair and gives a representative picture of a school's performance**. We realise that there are considerable concerns about the current way in which statutory assessment data is used including concern that the system is too 'high stakes', which can lead to unintended consequences such as over rehearsal and 'teaching to the test'.

We believe the system can be made fairer. Since much of the evidence and feedback we have received shows that the main concern lies with the way in which school accountability data is used, we suggest that any changes must begin with the accountability system.

As a result of the widespread view (held by many respondents) that information has been used unfairly in the past, some will always argue against any publication of statutory assessment data. However, **we believe that transparency is important and that the publication of school performance data provides useful information**. Nonetheless, we want to ensure that **published data are more comprehensive, and should only ever be viewed as a part of a much bigger picture**. We believe more information should be published to help parents and others to hold schools to account in a fairer, more rounded way.

We believe that accountability to parents is extremely important. There is widespread agreement about the importance of providing parents with good quality information both about their child and their child's school. We believe that information provided to parents should be accurate, easy to interpret and appropriately detailed so that they can support their children as effectively as possible.

We believe that more trust needs to be placed in teachers and there need to be credible, consistent national standards. Teachers are undoubtedly best placed to give an assessment of each individual child, and are uniquely qualified to identify the areas on which a child needs to focus in order to improve. **We would like to see a greater emphasis on teacher assessment within statutory assessment**, and summative teacher assessment to be given greater weight within the accountability system.

We wish to note that this Review has focused only on statutory summative assessment and has quite deliberately not touched on non-statutory assessment. We acknowledge that good teaching is wholly dependent on good assessment at every point, including throughout every lesson. Evidence highlights the importance of assessment for learning. We want to give schools and teachers as much time, energy and space as possible to use a continual repertoire of assessment techniques as they see fit.

However, we recognise that there are links between statutory summative assessment and ongoing formative assessment. We realise that some of our recommendations will depend on schools having formative assessment arrangements in place, for example effective tracking of pupils, which will support teacher assessment at the end of Key Stage 2. It will be important for the Government to consider fully what implications the changes to statutory assessment that we are recommending have for non-statutory formative teacher assessment.

While we believe strongly that a greater emphasis should be placed on summative teacher assessment, we also believe there is a continuing need for some assessment outcomes derived from externally-marked tests. It is inevitable that, if marking is done and moderated locally, there will be some variation and standards will not always be applied uniformly across the country. External marking gives greater confidence that mark schemes are applied fairly and consistently, which is crucial to the reliability of the results.

We believe that both summative teacher assessment and external testing are important forms of statutory assessment and have a valuable role to play. Both have strengths – and both have limitations. We do not believe that statutory assessment needs to rely only on one or other of these forms. We have therefore approached each aspect of statutory assessment from the perspective of what is educationally appropriate. This is the case when considering areas such as timing, computer based testing and the principle of 'testing when ready' in the fourth chapter of this Final Report, and is most evident in the third chapter when we make recommendations about each core subject.

Lastly, while it has been tempting in this debate to look backwards and focus on trends in the past or previous changes to the system, we have tried to be as forward-looking as possible. We see our recommendations as the way to deliver a stable system. We recognise that the new National Curriculum may require further changes to the statutory assessment system, but we believe we have set out long-term principles which should apply to the statutory assessment of the new National Curriculum as well as the current one.

We have set out both short-term recommendations and long-term principles that we think are important and which we believe will help deliver a fair and effective system, not only in the immediate future but for many years to come. We believe this stability is particularly important for the statutory assessment system which needs time to be properly implemented and developed.

Our recommendations

This summary outlines the recommendations which we believe will have the greatest impact and are of direct relevance to a wide range of schools, head teachers, teachers, parents and pupils. All the recommendations are explained in detail in the report itself. These recommendations follow on from the key principles and themes which we identified in the previous section. We have also made a number of suggestions for how these recommendations might be implemented in the short and long term, which are explained in detail in the final report.

Main uses of statutory assessment data

We recommend that **there should be only three main uses of the data from Key Stage 2 statutory assessment**, in addition to the purpose defined in primary legislation ("*to ascertain what pupils have achieved in relation to the attainment targets for that stage*"):

- holding schools accountable for the standard of attainment and progress made by their pupils and groups of pupils;

- informing parents and secondary schools about the performance of individual pupils;
- enabling benchmarking between schools; as well as monitoring performance locally and nationally.

We acknowledge that it is inevitable that statutory assessment data will also be used for a range of secondary uses. We would like to be clear that these are not the principal uses for which the system has been designed.

A greater focus on progress

We recommend that there should be a greater emphasis on pupils' progress within the accountability system. **We recommend that progress should be one of the two headline published measures (in addition to attainment), and any overall judgement of a school by the Government, local authorities or Ofsted should give at least as much weighting to progress as attainment.** At pupil level there should still be a minimum level of expected attainment, but we understand that some schools have to work much harder to get pupils to this level. We believe this should be recognised through a greater focus on progress which will take into account the different starting points.

The greater emphasis on progress should apply at individual pupil level as well as school level. We believe there should be a strong focus on the progress of every pupil.

We want schools that work hard to maximise the progress of pupils with low prior attainment to be recognised for doing so, which we believe is not always the case at the moment. **We therefore welcome the Government's commitment to introduce an additional published indicator of progress focusing on the lowest attaining pupils.** We believe this additional measure will help ensure schools focus on maximising the progress of every child, and will make it less likely that any schools focus on pupils at the level 3/level 4 borderline, to the detriment of other pupils.

We believe the statutory assessment system must support a culture of high expectations for all pupils and put the right incentives in place to ensure that the achievement of each pupil is maximised. We are aware that ensuring the progress and attainment of pupils with Special Educational Needs is maximised is an extremely complex challenge, currently being addressed in depth through the proposals which are being consulted on through the Green Paper, *Support and aspiration: a new approach to special educational needs and disability*[1]*.* We recommend that the Government should consider the outcomes of the consultation with a particular focus on ensuring that the achievement of all pupils with Special Educational Needs is appropriately recognised and celebrated within the accountability system. This needs to be the case both in mainstream schools and in special schools.

We believe it is crucial that the most able pupils are challenged effectively and that the system is able to recognise and celebrate their progress. We recommend that the Government should continue to provide level 6 National Curriculum Tests for schools to use on an optional basis, whose results should be reported to parents and secondary schools. If, following the review of the National Curriculum, any changes are made to the current system of levels, alternative arrangements should be put in place to ensure the most able pupils are challenged.

In order for progress to be given more weighting, a robust and consistent baseline is essential. **We recommend the moderation process at Key Stage 1 is developed further to be more consistently rigorous**. We suggest moderation at Key Stage 1 is better targeted so that schools where attainment and progress at Key Stages 1

[1] Department for Education, *Support and aspiration: A new approach to special educational needs and disability*, (2011).

and 2 are inconsistent are prioritised and moderated more frequently. We realise that many local authorities already target their moderation very carefully and we believe this should consistently be the case and made a formal requirement.

Broader accountability measures

We acknowledge the feedback from many respondents about league tables, which suggests they are a crude way of ranking schools, which is both unfair and unhelpful to anyone seeking an accurate comparison of different schools. **We believe that a greater range of published information will reduce the likelihood that league tables will be created focused on one indicator alone**. We believe a judgement about whether one school is better than another, based on just one measure, will simply not be credible when so much more information is available.

We recommend the introduction of three-year rolling averages, to be published alongside annual data, which we believe would take into account the volatility of results of individual cohorts and provide a sense of achievement over time.

We appreciate that schools are currently held accountable for pupils who arrive late in a Key Stage, and so a school's contribution to a pupil's learning is sometimes only partly reflected. **We recommend the introduction of additional attainment and progress measures for pupils who have completed the whole of Years 5 and 6 within a school.**

We believe that school-level measures across English as a whole subject are too broad to give a full picture of a school's performance in English. We recommend that schools' statutory assessment results in reading and writing should be published separately to allow schools to present a more rounded picture of their performance in English.

We acknowledge the concerns of some respondents who feel that tests are currently too concentrated into a single week, and the problems which arise when a pupil is absent on test day. We realise that pupils who are absent for a valid reason can currently take a National Curriculum Test within two school days. However, we do not believe an extension of two days goes far enough to resolve this problem. **We see the benefits to both schools and pupils of allowing a pupil who is absent on the day of the tests to take tests within an extended time frame** (not exceeding one week) and recommend that the Department considers trialling such a scheme in 2012.

Enabling effective benchmarking of schools

We believe that effective benchmarking by school managers is essential and that facilitating it requires additional tools and analysis. We recommend the continued use of Raiseonline by school managers because of the detailed information it provides.

We believe that benchmarking is most appropriate when schools with similar circumstances and challenges are compared. However, we believe every school should also be able to compare itself against the local and national average. We welcome the Government's commitment to publish 'Families of Schools' data and we believe that schools and Ofsted should look to use this tool as they see fit.

Publication of summative teacher assessment judgements

As well as ensuring that teacher assessment judgements for each pupil are reported to parents and secondary schools, we believe that a school's summative teacher assessment also provides useful information which the accountability system should take into account. We recommend that teacher assessment results should continue to be published.

However, we are concerned by feedback which suggests that, because teacher assessment results are often submitted after test results are announced, some schools perceive them as carrying less weight and may invest less time and effort into them. **We recommend that schools should submit their teacher assessment judgements ahead of receiving any test results. We believe this would help put greater emphasis on teacher assessment within the accountability system.**

Reporting to parents and secondary schools

We welcome the planned changes to the performance tables to ensure that information comparing schools locally is provided in a clearer and more accessible format to parents, incorporating a range of measures. We believe these changes will make it easier for parents to find additional, filtered, and up-to-date school-level information to compare schools locally. This will enable parents and the public to hold schools accountable for their contribution to children's education.

We believe that the pupil-level information provided to parents should be improved and that this would be useful to secondary schools as well as parents. We believe that if pupil-level information is easier to interpret and more detailed, this will both help support the learning of all pupils as soon as they arrive in Year 7, and also give parents a better picture of areas their child needs to focus on to improve. We recommend that pupil-level data both across mathematics and English and on the component parts of each of these subjects (i.e. for each attainment target) should be provided to parents and secondary schools.

We acknowledge the criticism of National Curriculum levels we have heard, including that they are too broad, not consistent across Key Stages, not specific enough about a pupil's ability in any given subject and too open to interpretation. In the short term, we believe we need to retain levels as a means of measuring pupils' progress and attainment. Key Stage 1 continues to be reported by levels, and therefore to measure progress robustly, Key Stage 2 results should be reported in the same way. However, we believe the introduction of a new National Curriculum provides an opportunity to improve how we report from statutory assessment.

International comparisons

Given the wide variation in assessment systems throughout the world, we believe that **participation in the specifically-designed international comparison studies should continue** as it will provide the most effective method of comparing standards internationally. We welcome the proposal in the 2011 Education Bill to make participation in international comparison studies mandatory for those schools selected.

Statutory assessment

We recommend that the **statutory assessment system should include both external testing and teacher assessment**, as we recognise that both have advantages as well as limitations. We recognise that teacher assessment can focus in detail on an individual child over the whole of Year 6, identifying the areas on which a particular child needs to focus in order to improve. We also recognise that external tests can give confidence that each child's work is being assessed fairly and consistently against the same mark schemes and national standards, producing robust and reliable data, particularly when tests focus on answers which are 'right' or 'wrong'.

We believe that **summative teacher assessment and externally-marked testing should not be seen in opposition to each other or used to judge each other**. We believe that both draw upon different and very valuable evidence bases. Testing and teacher assessment are therefore both important forms of assessment, and statutory assessment does not need to rely on only one or other of these forms. We feel that

decisions on whether a particular aspect of a subject should be tested or teacher assessed should be determined by the educational merits of the specific case.

Reading

We recommend that **reading should continue to be subject to externally-marked testing**. However, we do recognise that there are some concerns with the current reading tests, and believe they should continue to be refined over time.

We believe that the most crucial aspects of reading at the end of Key Stage 2 are accuracy, fluency and comprehension. We recommend that, subject to the National Curriculum Review, these skills should be brought out more clearly in the results of future tests that assess the new National Curriculum.

Writing

We recommend that **writing composition should be subject only to summative teacher assessment**. We believe this will encourage pupils to develop and demonstrate a broad range of writing skills over the course of Year 6, while avoiding the perverse incentives of the current system. While maintaining national standards is vital, we do not believe that this should come at the expense of promoting creativity. We believe teachers should assess widely across a range of genres and writing styles, and that this approach should give children more opportunities than the current system does to write in a wide range of ways. We recommend that **teacher assessment of writing composition should be subject to external moderation**.

We recognise there are some elements of writing (in particular spelling, punctuation grammar and vocabulary) where there are clear ‘right or ‘wrong’ answers, which lend themselves to externally-marked testing. **We recommend that a test of these essential writing skills is developed**.

We therefore recommend that in future **writing should be assessed through a mixture of both testing and summative teacher assessment**. Due to its importance, we believe that teachers’ own professional assessments of writing composition should always form the greater part of the overall writing statutory assessment result.

Speaking and listening

We recognise that speaking and listening are critical to the teaching and learning process. Children must be able to articulate their ideas and understand what they have been taught if teachers are to assess what they know. We recommend that **teacher assessment of speaking and listening should continue** and it should continue to inform schools’ overall teacher assessment of English. We recommend that the National Curriculum Review should consider how best to reflect the importance of speaking and listening in the new curriculum.

Mathematics

We recommend that **mathematics should continue to be subject to externally-marked testing** as the evidence we have received shows that the mathematics test is widely respected.

We acknowledge the concerns that results in the mathematics tests should not be determined by ability in reading. We recommend that in the development of future tests, the amount of reading in the mathematics test should be kept under review, to ensure that less able readers are not unfairly disadvantaged. In addition, we believe that the current principle that questions should be placed in order of difficulty should be carefully adhered to in the development of future mathematics tests.

Science

We recommend that **science should continue to be subject to summative teacher assessment**. As the current statutory assessment arrangements in science are relatively new, we acknowledge that their effectiveness is not wholly clear, but we believe the arguments for removing the external tests in science were justified. We believe it is important that national standards should continue to be monitored alongside schools' teacher assessment. We therefore recommend that **sample testing in science should continue**.

In the long term we recommend that the Government should continue to seek feedback from schools and the science community as to the appropriateness and effectiveness of the current arrangements, particularly in view of changes to the curriculum. We recommend that the current arrangements should be looked at again following the National Curriculum Review to ensure they are educationally appropriate for the new science curriculum.

Cluster moderation to support professional development

We understand the value of groups of teachers from a range of schools (including secondary schools) meeting on regular basis to build a shared understanding of educational standards and to discuss their assessment of pupils' work. We would encourage schools to form clusters in this way to moderate teacher assessment judgements with the aim of learning from each other and developing the assessment skills of the teachers involved. Many schools already participate in such networks, and we feel that other schools could benefit from adopting this approach.

Transition to secondary school

Given that we are recommending more detailed reporting of information to secondary schools and earlier collection of statutory teacher assessment judgements, **we encourage secondary schools to make wider use of the pupil-level data available from Key Stage 2 statutory assessment to support transition of new Year 7 pupils**.

Given the greater focus on teacher assessment information, we believe there is potential in encouraging cross-phase moderation of Year 6 pupils' work. We believe Year 7 teachers should be involved in the moderation of teacher assessment judgements of Year 6 pupils' writing composition work in particular, perhaps as moderators themselves. We also recommend that the Government should consider what incentives can be put in place to encourage Year 7 teachers to join in moderation exercises with Year 6 teachers designed to support professional development.

Transition from Key Stage 1 to Key Stage 2

We believe the same principle of encouraging cross-phase moderation should apply to infant and junior schools, and **we encourage moderation of Key Stage 1 teacher assessment judgements involving both Year 2 and Year 3 teachers**. We believe this approach would complement our earlier recommendation for a more consistently rigorous moderation process at Key Stage 1. It will help ensure that Key Stage 1 pupil-level data is robust and that Year 3 teachers feel confident in making wide use of it to understand their new intake

Timing of tests

The evidence and feedback to the Review suggests that changing the timing of end of Key Stage 2 tests to the beginning of Year 7 is a feasible option, but we believe is not the best solution to the problems of the current system. We do not believe there is now a compelling reason for tests to be moved to Year 7.

While there are arguments for moving tests earlier or later in the summer term, we do not believe these arguments are compelling either and so we recommend that tests should remain at the same point in the school year.

On-screen marking

We believe the wide use of **on-screen marking should be considered for Key Stage 2 tests** in addition to the science sample tests for which it is currently used. We recommend that the Government should learn from the evidence from science sample tests and plan what further trialling is needed with the aim of moving to a full roll out of on-screen marking.

Computer-administered testing

We believe the potential of computer-administered testing is enormous, but it needs to be approached with caution. We believe it should be explored further and piloted with a view to exploring the possibility of introducing it in the long term. We recommend the same approach should be taken with computer adaptive testing, given the advantages associated with pupils being able to sit their own personalised test according to their level of attainment.

Testing when ready

While we appreciate the potential benefits of the Single Level Test approach, we are not convinced by the model that was piloted. We do not believe that moving to a 'testing when ready' approach is the best way of achieving the purposes of statutory assessment under the current National Curriculum. However we suggest that the principle of 'testing when ready' should be considered in the future following the National Curriculum Review. We believe the 'testing when ready' approach would fit better with a curriculum which identified a core element which each pupil should master, which could be assessed when they are ready, before they move on to more advanced learning in the curriculum which may then be assessed later at a fixed point. We believe this approach should be considered following the National Curriculum Review.

Chapter 1 – Purposes of Statutory Assessment

In order to design a system which works effectively and is fair, it is important first to define the system's purposes. We have heard a great deal of evidence revealing significant concerns about the purposes of statutory assessment and the main intended uses for the resulting data.

Almost all respondents have at some point questioned the purposes of statutory assessment within the current system. There seems to be widespread concern that there are too many purposes, which can often conflict with one another. 71% of respondents to the online call for evidence believe strongly that the current system does not achieve effectively what they perceive to be the most important purpose.

Dr. Newton's work in this area is perhaps the most well known[2]. He has identified at least 16 purposes for which statutory assessment has been widely used. He cautioned that the current test results are better suited to some of these purposes than others. He suggested that assessments should be designed with a clear purpose in mind and urged caution when making inferences which the data were not intended to support.

The House of Commons Children, Schools and Families Select Committee's 2008 report on Testing and Assessment[3] drew similar conclusions from a wide range of evidence and criticised the use of national testing for multiple purposes.

As a result of this evidence, one of the key early decisions for this Review was to define clearly the purposes of statutory assessment in primary education and what we want it to achieve in the future.

We believe that, in addition to the main statutory purpose of summative assessment ("*to ascertain what pupils have achieved in relation to the attainment targets for that stage*"[4]), the following principal uses of statutory end of Key Stage 2 assessment data should apply:

a. **Holding schools accountable for the attainment and progress made by their pupils and groups of pupils.**

We believe that statutory assessment data can present part of the picture of a school's performance and, when complemented by other quantitative and qualitative information, can help parents, the public and central and local government to understand schools' strengths and weaknesses and hold schools accountable. Statutory assessment should provide information that can be used to measure the attainment and progress of pupils, broken down by groups of pupils as well as across whole cohorts. As the information will be used within the accountability system, this has implications for how it should be produced. Such data consequently needs to be considerably more consistent and reliable than might be the case in a system where results are not used for school accountability.

b. **Informing parents and secondary schools about the performance of individual pupils.**

---

[2] Newton, P.E., 'Clarifying the purposes of educational assessment', in *Assessment in Education: Principles, Policy & Practice*, 14(2), 149-170 (2007); Newton, P.E., *Evaluating assessment systems, Paper 1: Submission to the Education and Skills Select Committee Inquiry into Testing and Assessment,* Qualifications and Curriculum Authority (2007).
[3] House of Commons Children, Schools and Families Select Committee, *Testing and Assessment: Third Report of Session 2007-2008*, TSO, HC-169-I (2008).
[4] *Education Act 2002*, section 76.

We believe it is crucial that parents are provided with detailed information about their child's attainment and progress, broken down not only into English, mathematics and science, but also the different elements of each subject, so they can understand how to help their child improve. This information needs to be easy to interpret and reported on a regular basis. Secondary schools also need to receive and make use of similarly detailed information in good time to plan the transition of new Year 7 pupils, and ensure every pupil is appropriately supported and given opportunities to be stretched and challenged from the beginning and throughout their secondary education.

c. **Enabling benchmarking between schools, as well as monitoring performance locally and nationally.**

Benchmarking plays an important role in supporting the evaluation and self-evaluation of a school's performance. We envisage schools comparing themselves and being compared against schools in a similar context, as they will be facing similar challenges. However, schools should also have the opportunity to compare themselves to the local and national average. Producing such local and national averages will also provide useful information about local and national trends, which should be monitored on a regular basis to help inform policy development and delivery. Where models of assessment change, a new baseline may need to be created in order to use the results to monitor standards over time.

In coming to this conclusion, we have considered a wide range of potential **secondary uses** to which assessment data can legitimately be applied. These include:

For schools:

- Ascertaining whether pupils are making sufficient progress against expectations.
- Identifying students whose learning differs significantly from their peers, and where appropriate, identifying appropriate catch up or diagnosing learning difficulties and eligibility for special educational provision.
- Grouping students based on aptitude or attainment, and tailoring teaching and learning accordingly.

For parents:

- Providing parents with part of the information to identify the most desirable school for their child.

For the Government:

- Monitoring whether the performance of schools is rising or falling in relation to national expectations.
- Helping to ensure that resources are allocated appropriately.
- Helping to identify where intervention is needed and what measures are most appropriate.
- Helping to evaluate the success of educational programmes or initiatives.
- Providing information to inform the standard-setting process for future national curriculum assessment and testing.

We expect that statutory assessment will be used for some or all of these secondary uses in addition to the primary uses. However, **we would like to be clear that,**

while the information from statutory assessment will still prove useful for these secondary uses, they are not the principal uses for which the system has been designed. This will have important implications for the confidence with which such inferences can be made.

In addition, we believe that it is important to ensure we can compare our education system with education systems internationally. We consider that this is best done through participation in international comparison studies alongside (but not as part of) the system of statutory assessment, and we cover this at the end of the chapter on accountability.

Chapter 2 – Accountability

In this debate, testing, assessment and accountability are inextricably linked. Many opinions expressed about testing and assessment actually refer more to the accountability system. In this chapter we focus only on the accountability system and how we believe the results of statutory assessment should be used within it, while the next chapter will consider statutory assessment itself. This chapter sets out recommendations for school accountability at the end of Key Stage 2, explains our reasoning, and begins to outline how the system could be changed.

It may be helpful to set out our understanding of the current national accountability system. Firstly, schools' results are published in national Achievement and Attainment Tables; these tables also include school size, numbers and percentages of pupils with SEN, year-on-year comparisons, contextual value added scores, progress measures based on percentages making expected progress, average points scores and absence records. The Achievement and Attainment Tables are frequently used by the media to create league tables ranking schools based on their test results.

Secondly, Ofsted conducts inspections to provide an evaluation of how well a school is performing, resulting in a written report which includes an overall judgement using one of four grades: outstanding, good, satisfactory or inadequate.

Thirdly, the Government defines minimum 'floor standards' of performance based on attainment and progress in National Curriculum Tests. It uses these 'floor standards' to identify schools which it considers to be underperforming based on attainment and pupils' progress, as determined by their end of Key Stage 2 test results.

Fourthly, a wide range of school performance data are captured in RAISEonline, a data management system not accessible to the general public; the information from the system is used by school managers and local authorities, as well as Ofsted, to compare a school's performance with that of other schools, including those with similar pupil populations.

There is considerable debate about how schools should be held accountable. However, the evidence submitted to the Review overwhelmingly accepts that, as public bodies, schools should be held accountable for the education of their pupils. This is an important point of principle which underpins our recommendations about accountability.

The impact of school accountability

Strong evidence shows that external school-level accountability is important in driving up standards and pupils' attainment and progress. The OECD has concluded that a 'high stakes' accountability system can raise pupil achievement in general and not just in those areas under scrutiny[5].

There is clear evidence that the most effective systems in the world seek to combine significant operational independence for schools with effective accountability[6]. The PISA surveys identify two key system factors which make the most significant positive difference to student achievement – school autonomy, balanced by strong external accountability[7]. OECD evidence also suggests strongly that the use of

[5] OECD, *PISA 2009 Results IV: What Makes a School Successful? Resources, Policies and Practices*, (2010).
[6] Department for Education, *Schools White Paper: The Case for Change*, (2010).
[7] OECD, *PISA 2009 Results IV: What Makes a School Successful?*

external assessment data is important if accountability mechanisms are to be accurate and effective[8].

A range of research evidence indicates that systems in which local authorities, schools, teachers and, ideally, pupils are held accountable for their academic performance are associated with increased pupil attainment. Professor Wiliam's recent study[9] concluded that the effect of such assessment regimes could be equivalent to an increase in the rate of learning of approximately 20% and suggested that 'high stakes' accountability systems are the most cost-effective method yet developed for raising achievement.

Over the last three decades, research by the International Congress for School Effectiveness and Improvement (ICSEI) has explored the impact of schools and teachers on pupil outcomes, comparing it to the influence of pupil background or prior attainment. Work by Professor Sammons suggests that schools vary in their effectiveness in promoting pupils' academic results[10] and that such differences in academic effectiveness of schools tend to matter more for disadvantaged pupils[11]. These conclusions were illustrated by the longitudinal Effective Pre-school and Primary Education (EPPE) 3-11 study, which concluded that "*children who had the benefit of attending a primary school identified, through the National assessments, as academically more effective had better outcomes at age 11 than children who attended a less academically effective primary school, taking account of other background influences… highly disadvantaged children show a greater benefit than less disadvantaged children if they attend a highly effective primary school*"[12]. The variation in school effectiveness provides a strong argument for robust school accountability. Regardless of pupils' background or prior attainment, effective schools can make a great difference to pupils' outcomes – while ineffective schools may let down their pupils. Professor Sammons argues that having a single external summative assessment allows school effectiveness to be monitored, measuring the variation in pupils' attainment across the country and between schools.

In his evidence to the Review, Sir Michael Barber argued that if schools (and Government) are to make decisions based on evidence, they need regular assessment and data which are comparable over time. This enables analysis of the performance of different groups of pupils, provided it is based on valid, reliable and consistent data[13].

We believe the evidence that external school-level accountability drives up pupils' attainment and progress is compelling. We believe that providing evidence for external school-level accountability must remain a main purpose of any future approach to testing, assessment and accountability at the end of Key Stage 2.

Concerns over the school accountability system

However, many respondents have expressed concerns about the current way that schools are held accountable. The main concern is that the current system is too 'high stakes' for schools due to the publication of test results and their use for holding

[8] OECD, *PISA 2006: Science Competencies for Tomorrow's World, vol. 1: Analysis*, (2007).
[9] Wiliam, D., 'Assessing achievement at the end of KS2', submission to the Review (2010).
[10] Sammons, P., *School effectiveness and equity: making connections*, CfBT (2007).
[11] Sammons, P., 'Equity and Educational Effectiveness', in Peterson, P., Baker, E., McGaw, B. (eds), *International Encyclopedia of Education* vol.5, 51-57, Oxford (2010); Sammons *et al*, 'Children's Cognitive Attainment and Progress in English Primary Schools During Key Stage 2: Investigating the potential continuing influences of pre-school education', in *Zeitschrift für Erziehungswissenschaft 10*, 179-198 (2008); Scheerens, J. and Bosker, R., *The Foundations of Educational Effectiveness*, Oxford (1997).
[12] Sammons *et al*, *Influences on Children's Cognitive and Social Development in Year 6*, EPPE, DCSF Research Brief (2008).
[13] Mourshed, M., Chijioke, C., Barber, M., *How the world's most improved school systems keep getting better*, McKinsey (2010).

schools accountable for their performance, which leads to unintended consequences such as over-rehearsal and 'teaching to the test'.

NUT, NAHT, and ATL believe that test results are subject to unnecessary and damaging over-interpretation, and consequently undermine pupil learning[14]. NASUWT accepts the value of National Curriculum Tests, but believes that the way in which the results are used causes problems. The National Governors' Association (NGA) expressed concern at how responsive the school system has become to accountability systems. ASCL believes league tables are driving the whole education system, leading to assessment for its own sake rather than to identify what pupils need to learn. Some academic experts have blamed the 'high stakes' nature of the system for the current concerns of narrowing the curriculum. For example, Professor Torrance warns that "*the more individual student achievement is tied to system accountability, the more accountability measures will dominate student experience*"[15].

The following quotations from the online call for evidence indicate concerns with the way in which schools are currently held to account:

> *"Schools do need to be accountable and children and parents should receive accurate information about their progress and attainment. Having tests helps inform this process but should not stand alone as the only measure and only way schools are judged."*
>
> *"Just publishing results does not allow the public to make effective comparison between schools. Although high levels of attainment are important, so are other areas of school life, such as care and guidance and curriculum enrichment, all of which are graded by Ofsted. Just publishing results can give a very distorted picture, for example a school's results may be lower because they are more inclusive and have additional pupils with SEN."*
>
> *"All schools have their own character, comparison purely on end of KS2 results does not present a true picture of the school and its achievements. A school is a sum of its parts and needs to be judged on all its attributes, not a snapshot in time on a few curriculum areas."*

The online call for evidence asked which aspects of the current system most needed to change. The strongest view (from 50% of those who responded to the question) argued against publishing school-level performance data which could be used to create league tables, or stopped short of suggesting this but strongly highlighted some of the perceived problems that league tables cause.

The level of concern about the current system shows that change is needed. Feedback suggests that many head teachers and teachers think that the way aspects of school performance are judged within the current accountability system is unfair, and in some cases their hard work and successes are not appropriately rewarded. This needs to be addressed.

We are aware that many teachers and head teachers feel that the current combination of statutory assessment and the school accountability system constrains schools, compelling them to over-focus on what is assessed. Many heads have told us in discussion that they 'need' to concentrate much of Year 6 teaching on preparation for National Curriculum Tests in order to prevent results dropping. We

---

[14] ATL, NAHT, NUT, *Common Ground on Assessment and Accountability in Primary School,* (2010).

[15] Torrance, H., 'Using assessment in education reform: Policy, Practice and Future Possibilities', in Daniels H., Lauder H., & Porter J. (eds), *Knowledge, Values and Educational Policy,* London, 218-36 (2009).

recognise that the accountability system to date may appear to have encouraged this kind of behaviour.

Evidence suggests that this need not be the case. Good academic outcomes do not have to come at the expense of narrowing the curriculum. There is evidence[16] to suggest that very good primary schools are successful both in promoting pupils' academic progress (with most children attaining benchmark attainment levels at the end of Key Stage 2) and in ensuring high quality educational experiences across the curriculum.

We recognise that substantial changes to the school accountability system are already under way. The Government's recent Schools White Paper suggests a move away from centrally-driven targets, so that schools become more directly accountable to parents and the public[17]. For example, schools will no longer be required to set performance targets or be challenged by School Improvement Partners. Instead, the Government intends schools to be accountable to the pupils, parents and communities they serve through a broad range of published performance data.

Given the importance of external school-level accountability, we believe publishing data and being transparent about school performance is the right approach. We therefore want to retain external school-level accountability and support the availability of school-level performance data.

However, we want to ensure that, while each school is accountable for its performance, a broader range of indicators is available, including contextual information. **We believe that a broader range of published data would help ensure schools are held accountable in a fair way and would allow parents and others to focus on the measures and areas that are of most interest to them**. We accept that, however much data is published, there will always be a temptation to rank schools by one single indicator, but we believe it would not be credible or legitimate to suggest a school is more effective than another school based on one measure alone; such conclusions should always be based on a range of measures.

[16] Ofsted, *Twenty outstanding primary schools: excelling against the odds*, HMSO (2009).
[17] Department for Education, *The Importance of Teaching: the Schools White Paper 2010,* (2010).

Increased focus on progress

We believe that, in the past, the school accountability system has been too focused on performance based on attainment alone.

We believe that the Government is right to expect high standards of attainment for each pupil and to challenge and support every school to deliver this. It is important for the attainment of pupils to be monitored, so that schools where pupils have low attainment can be identified and given additional support. However, this should not be the sole focus of the school accountability system.

While high expectations of attainment are of great importance to pupils, measuring the progress pupils make better reflects the school's contribution to their education. Simply reporting pupils' attainment fails to acknowledge the amount of effort a school has put in and the difference in effort different schools need to make to get pupils to the same standards of attainment. Two schools may achieve the same attainment results, but one drawing pupils from an affluent area with pupils achieving higher levels on entry to the school may not have helped their pupils make as much progress as the one serving a disadvantaged community.

57% of respondents to the online call for evidence felt that the accountability system would be improved by informing parents about their children's progress and not just their attainment. The following quotations from the online call for evidence indicate the support for a greater emphasis on progress:

> *"Surely, progress from starting at primary school to leaving it is the most important. Some children will attain level 3 but will have come a long way to get that far. Others attain level 4 but have made slow progress."*
>
> *"The tests measure attainment rather than achievement/progress, so disadvantage schools in challenging circumstances in terms of providing information to parents and the public."*
>
> *"The most important purpose is to inform pupils about their progress, and parents about their children's progress. All that we do in schools should be about helping children to achieve their best, and working with parents is an important part of achieving this."*

We recommend that the school accountability system should focus on both attainment and progress. Attainment and progress should be the two headline published measures, and any overall judgement of a school by the Government, local authorities or Ofsted should give at least as much weighting to progress as attainment. We welcome the inclusion of a progress measure in the Government's new floor standards (which for primary schools is currently defined as fewer than 60% of pupils attaining level 4 in both English and mathematics and fewer than average making at least two levels of progress in English and mathematics).

We believe that this change has the potential to address much of the concern schools legitimately have that the accountability system is not fair. It is important for pupils both to have good attainment and to make good progress. For the accountability system to concentrate on either alone risks creating perverse incentives. Combined, the two measures will celebrate pupils' achievement and schools' contribution to it.

Ensuring a focus on the progress of all pupils

A greater emphasis on progress should apply at individual pupil level as well as school level. We believe there should be a strong focus on the progress of every pupil.

We believe that the progress of pupils with low prior attainment should be identified clearly within the accountability system. This is especially important given that the current pupil progress measure is a universal measure which is not designed to encourage schools to focus specifically on those pupils who have entered the Key Stage behind the nationally expected level. We want schools that work hard to maximise the progress of pupils who start from a low base to be recognised for doing so, which we believe is not always the case at the moment.

We therefore welcome the Government's commitment to introduce an additional published indicator of progress focusing on the lowest attaining pupils. We believe this additional measure will help ensure schools focus on maximising the progress of every child, and will make it less likely that any schools focus on pupils at the level 3/level 4 borderline to the detriment of other pupils. It will help give schools that do a good job at accelerating the progress of the pupils with low prior attainment credit for the work that they do.

We are aware that it is possible to go even further in recognising the progress of low attaining pupils. For example the State of California's 'Academic Performance Index'[18] ensures that progress for the lowest performing pupils has a greater impact on a school's score than the progress of higher performing pupils does. We are not suggesting this approach should necessarily be implemented here. However, in addition to the introduction of the new indicator, we suggest that in the future the Government considers what more could be done to encourage schools to focus on the progress of the lowest attaining pupils.

We have heard feedback from head teachers that the current accountability system may discourage schools from admitting pupils identified with Special Educational Needs because they are concerned that it could have a detrimental impact on their overall results. We share this concern, but we are also aware of potential perverse incentives associated with removing pupils from the conventional measures. We believe the statutory assessment system must support a culture of high expectations for all pupils and put the right incentives in place to ensure that the progress of each pupil is maximised. We did consider removing the pupils working below National Curriculum levels from the conventional accountability measures to reduce the risk of perverse incentives, but as this only includes 0.6% of pupils we do not believe this change would appropriately address the concern.

We are aware that ensuring the progress and attainment of pupils with Special Educational Needs is maximised is an extremely complex challenge, currently being addressed in depth through the proposals which are being consulted on through the Green Paper, *Support and aspiration: a new approach to special educational needs and disability*[19]*.* We do not believe this challenge can be tackled effectively through the statutory assessment system in isolation, and we therefore do not wish to make any recommendations concerning pupils with Special Educational Needs, since we feel they should be considered alongside a wider package of changes. **However, we recommend that the Government should consider the outcomes of its consultation with a particular focus on ensuring that the achievement of all pupils with Special Educational Needs is appropriately recognised and celebrated within the accountability system. This needs to be the case both in mainstream schools and in special schools.**

We recognise that the current system of National Curriculum tests can appear to place a ceiling on attainment for the most able pupils. This has important implications for measures of progress, since a pupil who achieves level 3 at the end of Key Stage 1 can currently only achieve level 5 in the end of Key Stage 2 tests, and can therefore only make two levels of progress (currently the expected rate of progress). Allowing pupils to attain level 6 at the end of Key Stage 2 would enable pupils with

[18] Further detail can be found at www.cde.ca.gov/ta/ac/ap.
[19] Department for Education, *Support and aspiration: A new approach to special educational needs and disability*, (2011).

high Key Stage 1 attainment to make better than expected progress. Secondary schools receiving pupils who had attained level 6 would understand that these pupils would need to be particularly challenged and stretched from the start of Year 7. The following quotation explains how some respondents feel about this issue:

> *"Allow pupils to achieve more than Level 5. The ceiling at Level 5 restricts pupils and also limits the school's ability to demonstrate Value Added and pupil progress. Many of our pupils achieve Level 3 in KS1 and therefore cannot show value added in KS2 as they can only get a Level 5."*

It is important to challenge the most able pupils. We welcome the Government's decision to make level 6 tests available to schools on an optional basis this year. We believe that these optional tests could allow particularly able pupils an opportunity to develop and fully demonstrate their knowledge and understanding.

However, we do have some concerns, in particular over the extent to which it will be possible for primary schools to cover enough of the Key Stage 3 curriculum to allow pupils to attain level 6. NFER, one of the few respondents who commented on this issue, suggested that it would be more appropriate to award a 'high 5' than a level 6.

We believe that the Government should continue to provide level 6 National Curriculum Tests for schools to use on an optional basis, whose results should be reported to parents and secondary schools. If, following the review of the National Curriculum, any changes are made to the current system of levels, alternative arrangements should be put in place to ensure the most able pupils are challenged.

Key Stage 1: baseline to measure progress

A greater focus on progress will make it even more important that there is a robust baseline at Key Stage 1. At present schools are required to administer internally-marked statutory tests and tasks during the course of Year 2, which should inform the teacher assessment judgements reported at the end of the year. Local authorities are required to moderate summative teacher assessment in at least 25% of their schools.

Since Key Stage 1 summative teacher assessment results are not published at school level, these arrangements seem appropriate. Many heads and teachers have expressed support for this approach. However, we have also heard some feedback from heads (particularly of junior schools) that teacher assessment judgements in infant schools may be 'inflated'.

> "*KS2 teachers often feel the KS1 results are over assessed which leads to impossible targets for KS2 teachers. This is a particular problem in areas with middle schools… If KS1 and KS2 are in the same school it is less likely that a teacher would over assess a pupil as they would soon be found out by their colleagues. We have complained for years about this problem but nothing has ever been done*."

> "*The most flawed part* [of Key Stage 2 tests] *is the measure of progress. We find ourselves in a situation where our main feeder Infant School is sending us KS1 results which can be at literally twice the national rates for L3 which we are then judged against*."

We have also heard evidence that some all-through primary schools may under-assess pupils – particularly those on the borderline between levels 2 and 3 – in order to improve their chances of getting higher rates of progress between Year 2 and Year 6.

Evidence suggests that Key Stage 1 attainment results are higher on average in infant schools than in all-through primary schools. However, adjusting for the

characteristics of pupils in infant and primary schools, including factors such as eligibility for free school meals (FSM), accounts for a substantial part of the difference. After contextualising, infant schools still have higher attainment than primary schools, but the attainment gap is considerably smaller. Differences in pupil context in these types of schools account for a large part of the attainment gap.

It is worth noting though that between 2004 (the last year where end of Key Stage 1 tests were the main form of accountability) and 2010 the attainment gap has widened. The gap in percentage points of pupils between infant and primary schools attaining level 2 or higher in reading, writing and mathematics increased by 1.5, 2.4 and 1.5 respectively. The attainment gap has therefore become greater under the current system (based on summative teacher assessment informed by tests and tasks) than the previous system (based on external end of Key Stage 1 tests).

As a result of the greater focus on progress and, given that Key Stage 1 results count towards the baseline for progress measures, **we recommend the moderation process at Key Stage 1 is developed further to be more consistently rigorous. We suggest moderation at Key Stage 1 is better targeted so that schools where attainment and progress at Key Stages 1 and 2 are inconsistent are prioritised and moderated more frequently**. **We realise that many local authorities already target their moderation very carefully and we believe this should consistently be the case and made a formal requirement.**

The next chapter goes on to outline our recommendations for changes to Key Stage 2 statutory assessment. We believe that to measure progress robustly there needs to be a clear link between Key Stage 1 statutory assessment and Key Stage 2 statutory assessment. **We recommend that, in the long term, the Government should ensure that Key Stage 1 statutory assessment reflects changes at Key Stage 2 and the introduction of a new National Curriculum.**

Ofsted inspections

We have heard a great deal about the importance of Ofsted inspections and the value of their judgements and reports within the accountability system. 61% of respondents to the online call for evidence thought that Ofsted's inspection reports should be used to compare schools and hold them accountable. Ofsted's inspection process takes into account a school's self-evaluation, the previous Ofsted report, conversations with the head teacher, pupils, staff and governors, lesson observations, and the views of parents, in addition to the school's assessment data. Such a comprehensive professional evaluation will always give a more rounded picture of a school's performance than assessment data alone. We want Ofsted inspections to continue to be reliable and credible; we have therefore considered the feedback to the Review about Ofsted inspections, the process behind them and the data they use.

The following quotations give a sense of how much many online call for evidence respondents value Ofsted inspection judgements and reports, a point which was raised by a range of head teachers, including those who work in schools in challenging circumstances:

> *"Ofsted give an accurate view of a school's performance, they provide a more rounded picture than a set of figures ever can."*
>
> *"The Ofsted grading gives a fairer way of comparing schools as this explores more areas."*
>
> *"Much of the information that parents require has been regularly provided by Ofsted through an increasingly rigorous process of short notice national school inspections."*
>
> *"The comparison of achievement across a school should be made through Ofsted, which should compare the progress made by children in a school; not the percentage reaching a certain level."*

There is a perception that that Ofsted inspectors attach too much significance to a school's recent test results. 15% of respondents to the online call for evidence commented on the impact of test results on Ofsted's inspection judgements, some expressing deep concern. NUT, NAHT and ATL believe that Ofsted places an undue emphasis on test results as a proxy for school effectiveness. They argue that the inspection system should take greater account of the whole school's achievements. The House of Commons Select Committee's 2010 Report on school accountability concluded "*when evaluating academic attainment, we recommend that Ofsted gives less evidential weight to test results and derivative measures and gives more weight to the quality of teaching and learning observed by inspectors in the classroom*"[20]. Ofsted's response to the Committee confirmed that such measures had already been introduced in its 2009 inspection framework[21].

Having considered how Ofsted uses data, we concluded that the current process uses test results to inform rather than determine the overall grading of a school. In addition, when assessing the results, inspectors consider not only the current year's results (including both test results and reported teacher assessment judgements) but also the pattern over the last three years, as well as the latest information held by the school (for example, results from optional National Curriculum Tests).

---

[20] House of Commons, Children, Schools and Families Committee, *School Accountability: First Report of Session 2009-10*, HC88-1 (2010).

[21] House of Commons, Children, Schools and Families Committee, *School Accountability: Responses from the Government and Ofsted to the First Report of the Committee, Session 2009-10*, HC 486 (2010).

We welcome the fact that, under the current framework[22], it is pupil achievement rather than attainment which is given most weight. The judgement on 'achievement' takes account of both attainment and pupil progress. The assessment of pupil progress takes account of how well pupils make progress relative to their starting points, including whether there is any significant variation between groups of pupils, and how well pupils with Special Educational Needs and/or disabilities make progress.

There is a very strong correlation between Ofsted's judgements on achievement and overall effectiveness, with 93% of schools inspected in 2009/10 having the same judgement for both[23]. This contrasts with 55% having the same judgement for attainment and overall effectiveness. In over a third of schools, the overall effectiveness judgement was higher than the attainment judgement, reflecting the importance of the progress pupils make. We therefore believe it is highly likely that where achievement is strong and attainment is less strong, this will be reflected in a positive judgement for a school's overall effectiveness. More generally, our discussions with Ofsted confirm that inspectors already consider the full range of evidence available in reaching their judgement of a school's performance.

In its consultation document, *Inspection 2012*[24], Ofsted set out its initial proposals for judging pupil achievement in future. This includes giving particular attention to how well pupils learn, the quality of their work and the progress they have made since joining the school, as well as pupils' attainment by the time they leave school.

We recognise that Ofsted currently takes account of more that just test data in forming its judgments. We welcome the proposal to place greater emphasis on pupils' progress in the inspection process.

[22] Ofsted, *The framework for school inspection in England under section 5 of the Education Act 2005, from September 2009*, (2009).
[23] Ofsted, *The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2009/10*, HMSO (2010).
[24] Ofsted, *Inspection 2012: proposals for inspection arrangements for maintained schools and academies from January 2012,* (2011).

Broader accountability measures

As well as an increased emphasis on progress to give a more representative picture of a school's contribution to pupils' learning, additional accountability measures should be introduced to allow schools to draw on a wider range of published data to present a picture of their performance each year. This approach will also allow parents and the public a wider choice of indicators based on what is most important to them.

We understand that results each year can vary because of factors other than the quality of teaching and learning in a school, including variations in cohorts and pupil mobility. This is particularly true in small schools, where an individual pupil's results can have a disproportionately large effect on the school average. We believe that introducing new measures to reflect these factors will give schools additional ways to present and celebrate their performance and achievements. It will also render comparisons between schools based on league tables which use just one set of figures, much less legitimate or credible. Headline measures of attainment and progress will still remain important, but we want head teachers and teachers to be able to present a richer and more representative picture of their school each year.

The following quotations demonstrate the way many of the online call for evidence respondents feel about the current range of measures:

> *"Schools should be accountable for their performance but the system must be a fairer one."*
>
> *"I had a child join my school 2 weeks before SATs – she was a school refuser and came from a next door school. She achieved level 2 and was then included in our results and omitted from the next door school's results. This reduced our success on paper. This is an unfair reflection of the huge amount of work that has gone in to improving the life chances of the children in my school."*
>
> *"Using test results as a way to compare schools is unfair as it does not take into consideration other factors, most importantly inclusion, the emotional literacy of a school or the ethos of the school."*

One of the most consistent criticisms of the current school accountability system is directed at the way in which published data is turned into league tables. When asked what parts of the current system most need to change or be improved, 50% of online call for evidence respondents said that league tables should be removed. Feedback from many respondents has suggested that they are seen as a very crude way of ranking schools, which is both unfair and unhelpful to anyone seeking an accurate comparison of different schools.

We believe that a greater range of published information will reduce the likelihood that league tables will be created focusing on one indicator alone, which are unlikely to give a true comparison of schools' performance. We believe a judgement about whether one school is more effective than another, based on just one measure, will simply not be credible when so much more information is available.

Rolling averages

As we have noted above, the volatility in individual pupils' outcomes can have an effect on school results. This is exacerbated for small schools. In addition, some of the additional measures we propose may report on fewer than ten pupils in a number of schools. Currently the Government does not publish data for fewer then ten pupils, as the cohort may be too small to avoid identifying individual pupils. Potentially this reduces the number of schools for which data could be published.

We recommend that the main published statutory assessment data should be presented with rolling averages as well as annual data. This would take into account the volatility of results of individual cohorts and provide a sense of achievement over time. We believe that, because rolling averages take account of the results of a much larger cohort of pupils, they are particularly useful for small schools, where the size of each year's cohort means that average results will be more significantly influenced by the attainment of individual pupils.

Evidence shows that rolling averages would reduce the variability in schools' results from year to year. The following table compares the range of volatility in school results[25] between annual data and three-year rolling averages:

Cohort size	Average year-on-year change in results (% attaining level 4+ in English and mathematics)			
	English		Mathematics	
	Annual results	Three-year rolling averages	Annual results	Three-year rolling averages
All schools	9.30%	3.31%	9.69%	3.63%
Fewer than 16 pupils	13.20%	4.92%	14.34%	5.33%
16-30 pupils	9.75%	3.43%	10.14%	3.75%
31-60 pupils	8.12%	2.85%	8.26%	3.17%
More than 60 pupils	6.26%	2.22%	6.36%	2.46%

As the table shows, the year-on-year change in annual results is largest for smaller schools. Employing a rolling average reduces the year-on-year change by almost two thirds. We believe this would make a considerable difference to ensuring the performance of each school is more fairly represented and not excessively skewed or distorted by a particular cohort or the inclusion of particular pupils in that cohort. We believe a rolling average over three years would give a more rounded picture of a school's performance over time. Feedback has suggested this would be of greater relevance to parents, who are most concerned with the overall trend of a school's performance. The results in any one year may not reflect the school's performance when their child reaches the end of Key Stage 2. These new rolling averages may need to be carefully presented and explained to parents who are currently used to annual data.

We recommend that rolling averages should be over three years, as data analysis suggests that there is little reduction in volatility when moving from 3 to 4 or 5 year rolling averages, even for small schools.

[25] Average year-on-year change in percentage of pupil results (percentage achieving level 4+ in English and mathematics National Curriculum Tests) for annual results and three-year rolling averages for different cohort sizes from 2005 to 2010. The table shows absolute changes, i.e. a change of -3% would count as 3%.

Pupil mobility

We acknowledge the feedback to the Review which suggests that moving between schools can be disruptive and can adversely affect a pupil's progress and attainment[26]. The Government defines a 'mobile' pupil as one who joins a school at any point after the September of the academic year prior to assessment. Therefore, for Key Stage 2 statutory assessment, a 'mobile' pupil joined the school after the September of Year 5. In 2010 10% of the Year 6 cohort who sat National Curriculum Tests had moved school during Year 5 and 3% had moved during Year 6.

There is a clear association between mobility and Key Stage 2 attainment and progress. In 2010 nationally around 74% of pupils attained level 4+ in English and mathematics at the end of Key Stage 2. This compares with 75% for pupils who were not 'mobile', falling to 68% for those who moved school during Year 5 and falling further to 62% for Year 6 movers. Similarly, the proportion of pupils making or exceeding the expected levels of progress from Key Stage 1 to Key Stage 2 is lower for 'mobile' pupils than their peers. Even when pupils have achieved or exceeded expectations at Key Stage 1, fewer pupils make the expected levels of progress by the end of Key Stage 2 if they have changed schools in Year 5 or Year 6. For example, of the non-mobile pupils who achieved level 3 at English at Key Stage 1, 76% achieved 2 or more levels of progress, and this compares to 68% for Year 5 movers and 67% for Year 6 movers. The same pattern is evident for mathematics where 84% of non-mobile pupils who achieved level 3 at Key Stage 1 made the expected level of progress by the end of Key Stage 2, but only 75% of movers in Year 5 and 76% of movers in Year 6 did so.

We appreciate that schools are held accountable for pupils who arrive late in a Key Stage, and so the results of those pupils only partly reflect the contribution to their learning that their new schools make. We do not believe this is always fair on schools, particularly those with relatively high proportions of 'mobile' pupils.

We recommend the introduction of additional attainment and progress measures for pupils who have completed the whole of Years 5 and 6 within the school. However, we do understand this may not always be possible in small schools or schools where fewer than 10 pupils have left or arrived, as the cohort could be too small to avoid the identification either of the pupils who have completed the entire Key Stage, or of those who have not. We hope that also providing a three-year rolling average may make this data available for more schools.

We understand that mobility is a complex issue, and within the category of pupils who move school during Year 5 or Year 6, some move school more than once. We recognise also that 'mobile' pupils can come from some of the most disadvantaged communities, and we also need to ensure that there are no perverse incentives discouraging schools from meeting their particular needs. **We recommend that data on the 'mobility' of pupils who have joined in Years 5 and 6 should be published.** This data should best reflect the mobility of pupils in each cohort, reflecting both the proportion of pupils who are 'mobile' and how often those pupils move school.

School-level measures in reading and writing

We believe that measures across English as a whole subject are too broad to give a full picture of a school's performance. It would be much more helpful to publish separate measures of reading and writing. **We recommend that schools' statutory assessment results in reading and writing should be published separately, to**

[26] Hattie, J., *Visible Learning: a synthesis of over 800 meta-analyses relating to achievement*, London, (2008). Professor Hattie identifies mobility between schools as the single most negative influence on pupil achievement.

allow schools to present a more rounded picture of their performance in English.

However, we recognise that a composite English level may be necessary, in particular to provide a baseline against which to measure progress made at secondary school.

If an overall English measure is necessary, we suggest that options are explored for how it could be generated from a combination of the results of the reading test, teacher assessment of writing composition, and the proposed new test of spelling, punctuation, grammar and vocabulary described in the next chapter.

We recognise that schools currently report an overall teacher assessment level in English which reflects the teacher assessment judgements for speaking and listening, reading and writing. We think this overall teacher assessment level for English should continue to be reported and published.

Additional measures and contextual information

We welcome the Government's decision to publish other measures to help to give a more rounded picture of a school's performance. We suggest that additional contextual information could also include the proportion of pupils eligible for free school meals, or the proportion of those pupils eligible for the pupil premium in each Year 6 cohort. Schools will therefore be able to refer to a greater range of published data when explaining their pupils' performance.

We believe these additional measures will help schools give a more representative picture of their performance.

Allowing absent pupils to take tests within a given time frame

We acknowledge the concerns of those who feel that tests are currently too concentrated in to a single week, and the problems which arise when a pupil is absent on test day.

The quotes below give a sense of the frustration expressed by online call for evidence respondents about this issue:

> *"Currently if a child is absent they get no score. This is unfair to the school and the child."*
>
> *"If a pupil is ill and misses the tests, there is no opportunity for that child to be tested and the results to count. If the tests are to be used for accountability, then all pupils' results should count. This can be particularly evident for small cohorts where 1 child can represent 5%. If that child is predicted to get a level 5 and they are absent, the school's results are dramatically reduced. This does not lead to fair comparisons between neighbouring schools and can affect parents' decisions about which school they choose for their children."*

We realise that pupils who are absent for a valid reason can currently take a National Curriculum Test within two school days. However, we do not believe an extension of two days goes far enough to resolve this problem. We see the benefits to both schools and pupils of allowing a pupil who is absent on the day of the test to take the test within an extended time-frame (not exceeding one week), subject to checks to ensure reasons for absence are genuine and the necessary security measures are in place. **We recommend that the Government trials such a scheme in 2012, examines the impact, and considers whether to make this a permanent change.**

The publication of summative teacher assessment judgements

We believe a school's summative teacher assessment provides useful information which the accountability system should take into account. As well as ensuring that teacher assessment judgements for each pupil are reported to parents and secondary schools, **we recommend that teacher assessment results in each school should continue to be published in Achievement and Attainment Tables, as has been the case since 2010.** We believe the publication of teacher assessment results helps to show that they are of fundamental importance. Teachers know each child as an individual, and can provide an assessment across the full range of the curriculum and over a whole year. This should be acknowledged in the accountability system.

The following quotations from the online call for evidence show the importance which the profession places on teacher assessment judgements:

> *"Teacher assessment needs to be recognised and given a higher weighting relative to test scores."*
>
> *"Tests are currently seen as the ultimate piece of data that is used for every possible reason. Teacher assessment should be used as part of the picture."*

However, feedback suggests that, because teacher assessment results are often submitted after test results are announced, some schools perceive them as carrying less weight and may invest less time and effort into them. At present the deadline for schools to submit teacher assessment results falls after the return of National Curriculum test results – though in both 2009 and 2010 just over a third of summative teacher assessment results had been submitted ahead of schools receiving test results in early July.

The analysis below shows that, where schools submit summative teacher assessment data before the release of test results, there is generally a greater difference between the percentages of pupils attaining the expected level. This effect was more pronounced in 2010 than in 2009, and there was a greater effect in mathematics than in English.

This analysis suggests that if all schools reported their summative teacher assessment judgements before they received their statutory test results, the results are likely to be different. Important research (notably by Professor Harlen) demonstrates that summative teacher assessment judgements can legitimately differ from statutory test results, since they are informed by different evidence bases[27]. We wish to be clear that any difference between summative teacher assessment judgements and statutory test results should not be seen as indicating that either result is 'wrong'. We believe both summative teacher assessment judgements and external test results give important information, drawn from different evidence. We want to ensure that summative teacher assessment judgements are given adequate time and attention, independent of external test results, so that they can have appropriate weighting.

The table below shows that the average difference between the test and teacher assessment results in percentage points was highest in 2009 for English for schools which submitted their teacher assessment results before the release of test results, and lowest in 2010 for mathematics for schools which submitted their teacher assessment results after the release of test results:

[27] Harlen, W., *A Systematic Review of Evidence of the Impact on Students, Teachers and the Curriculum of the Process of Using Assessment by Teachers for Summative Purposes in Research Evidence,* EPPI-Centre, Social Science Research Unit, Institute of Education (2004).

Average difference in percentage points between test and teacher assessment results		English	Mathematics
2010	Before release of test results	+/-6.27	+/-5.47
	After release of test results	+/-5.22	+/-4.55
2009	Before release of test results	+/-6.57	+/-5.83
	After release of test results	+/-5.84	+/-5.16

We recommend that schools should submit summative teacher assessment judgements ahead of receiving any test results, and that these summative teacher assessment judgements should be published. We believe this would put greater emphasis on teacher assessment.

We also believe that collecting teacher assessment results earlier would encourage primary schools to pass them up to secondary schools earlier, to inform planning for the Year 7 intake and to target teaching and learning more effectively. We believe this change, combined with our recommendations for more detailed reporting to secondary schools, will improve the quality and timing of the information available to secondary schools.

Reporting to parents

One of the main uses of statutory assessment data which we have identified is providing information to parents about the attainment and progress of their children. The evidence we have received shows that there is widespread agreement about the importance of providing good quality information to parents.

Research evidence suggests that parents of primary school age children value the information from external tests, and highlights other information that parents find useful. Survey respondents wanted information about the performance of primary schools to be available to the public, and valued National Curriculum Tests for providing information about the how their child and their child's school is performing.

The National Confederation of Parent Teacher Associations (NCPTA) survey in 2008[28] showed that a strong majority of parents (78%) placed a high or medium value on their children taking external tests. Only a minority of parents (21%) reported that they were of no use. A DCSF Ipsos MORI survey[29] of 939 parents in England with children of school age in March 2009 found that 75% of parents thought information on the performance of primary schools should be available to the general public. 70% of parents placed value on the tests in providing information about how their child's school is performing. 65% of parents placed importance on their child taking part in Key Stage 2 tests, while 44% thought that tests should stay as they are. 36% of parents wanted to replace tests, but of those, most felt that some form of testing or assessment should remain.

When the previous Government consulted on its School Report Card proposals, a DCSF survey[30] found that the majority of parents (91%) and the public (84%) said it was very important to them personally to know how well each school performs. The majority of parents (87%) and the public (82%) agreed a lot or a little that test and exam results are one important measure of a school's performance. In addition to test results, parents and the public would most like information on behaviour at each school (parents 65%/public 56%). The next most popular kinds of information were attendance rates and how well pupils progress during their time in the school.

In contrast, a 2009 NAHT survey[31] indicated that the majority of parents surveyed by the union did not favour external testing. The survey suggested that parents instead value information such as attendance rates and behaviour at their child's school or prospective schools.

Many respondents to the Review commented on the information parents find valuable and whether the current system provides it. 59% of online call for evidence respondents (predominately head teachers and teachers) said that more weight should be given to teacher assessment. They said that, rather than being provided with a 'snapshot' test result, parents would welcome the more comprehensive information that could be provided from teacher assessment. Respondents to the online call for evidence believed that accurate information based on teacher assessment should be given throughout the child's education and not just at the end of Key Stage 2. Many proposed that this information should be given with a clear indication of how individual children were progressing against age-related expectations.

---

[28] National Confederation of Parent Teacher Associations (NCPTA) survey, *The Parent Perspective: SATs and National League Tables*, (October 2008).
[29] DCSF Ipsos MORI, *Survey of 936 parents in England with children of school age,* (March 2009).
[30] Department for Children, Schools and Families, *21st Century Schools: A World-Class Education for Every Child; A School Report Card: consultation document. Analysis of responses to the consultation documents*, House of Commons (2009).
[31] NAHT, *Survey of 10,465 parents,* (January/February 2009).

A range of stakeholders have given their views about what information parents find useful. The National Governors' Association (NGA) reported that some parents like external accountability and believe Key Stage 2 tests are important. However parents sometimes apply additional pressure on their children to perform well, for example setting extra work to be undertaken at home, which some respondents felt may contribute to pupils being 'over-rehearsed'. NAHT said that "*parents, politicians and the public have a right to see information about the performance of their children and their schools. We just don't believe the raw data they get from the current league tables is relevant or helpful*"[32].

We believe the evidence and feedback shows that parents value academic information about both their child and the school as a whole, as well as additional information which gives a more rounded picture of the school.

We believe that the combination of the recommendations set out already in this chapter and the current requirements on schools to publish information (including recent changes as part of the Government's transparency agenda) will go a long way towards ensuring parents are provided with the school-level information they value. In addition to this quantitative information, Ofsted reports provide useful qualitative information on areas such as behaviour, curriculum, leadership and management, which cannot be measured or reflected by the production of data.

We welcome the planned changes to the performance tables to ensure that information comparing schools locally is provided in a clearer and more accessible format to parents, incorporating a range of measures. We believe these changes will make it easier for parents to find additional, filtered, and up to date school-level information to compare schools locally.

However, many respondents to the Review have argued that the pupil-level information currently provided to parents is difficult to interpret and lacks detail. **We believe the pupil-level information provided to parents should be improved, and that this would be useful to secondary schools as well as parents.**

Reporting pupil-level results to parents and secondary schools

We believe the current reporting of statutory assessment to parents does not provide sufficiently detailed information for parents and secondary schools to identify the specific areas where each pupil needs to improve. In addition, primary teachers and heads understandably feel frustrated that, having built up a detailed knowledge of their pupils' strengths and weaknesses, secondary schools frequently make little use of the resulting assessment data. While the breakdown of pupil marks underpinning test results offers much more detailed information, secondary schools rarely have the time to undertake detailed analysis for a large pupil cohort, often drawn from a range of feeder schools. In addition the current tests are designed to produce reliable aggregated information for each pupil and each sub-element of each test is not designed to be completely reliable in isolation.

We believe that pupil-level data both across each subject and on its component parts should be provided to parents and secondary schools. Information provided to parents and secondary schools at pupil-level should be considerably more detailed than the published school-level information. Schools are already required to determine teacher assessment levels in each attainment target in English, mathematics and science, and to submit an overall subject level. **We recommend that schools should be required to submit teacher assessment levels both for the overall subject and for its attainment targets (or any equivalent in the future), and that this data should be provided to secondary schools.**

[32] NAHT, 'A better place on assessment and accountability', submission to the Review, (2010).

We have heard widespread concern that secondary schools make limited use of the information they receive about their new intake. Many secondary school respondents have expressed concern that National Curriculum Test results or primary schools' teacher assessment are not always a suitable proxy for the attainment of pupils on entry to Year 7. **We believe that if pupil-level information is easier to interpret and more detailed, this will both help support the learning of all pupils as soon as they arrive in Year 7, and give parents a better picture of their children's strengths as well as the areas on which they need to focus in order to improve.**

Parental surveys

We believe that, as well as placing great emphasis on providing the right information to parents, the accountability system should take into account information parents themselves can provide about a school. We believe that parental surveys are an effective way of enabling parents to reflect their views. While we would expect each school to take responsibility for responding to such surveys, where consistent and significant concerns are expressed by parents at a school, we believe it would be appropriate for the accountability system to give them careful consideration. If necessary this should trigger greater scrutiny of a school.

We recognise that Ofsted already runs a parental survey system whereby schools are required to send parents a standard questionnaire when they are notified that they will be inspected. We believe this ensures that the views of parents are gathered at the point of inspection and enables their opinions to be taken into account in the inspection process.

We note Ofsted's proposal in its consultation on a new inspection framework[33] to engage with parents outside of the inspection process. **We welcome this proposal and also encourage schools to gather the views of parents regularly, as many schools already do.** We support this approach as we believe parental feedback is very valuable and adds to the picture given by results, data and information presented by schools themselves.

National Curriculum levels

Some respondents have criticised the current system of National Curriculum levels. Feedback suggests that they are too broad, not consistent across Key Stages, not specific enough about a pupil's attainment in any given subject and difficult to interpret, including for parents.

In the short term, we believe we need to retain levels as a means of measuring pupils' progress and attainment. Key Stage 1 continues to be reported by levels, and therefore to measure progress robustly Key Stage 2 results should be reported in the same way. We believe this is the case because it is important that progress is measured in a way that is meaningful to those who use the information and a change to levels in the short term is likely to put this at risk.

However, in the long term, **we believe the introduction of a new National Curriculum provides an opportunity to improve how we report from statutory assessment. We believe it is for the National Curriculum Review to determine the most appropriate way of defining the national standards which are used to categorise pupils' attainment.**

We realise that, in order to measure progress, it is necessary to have an appropriate scale against which attainment and progress can be measured at various points. For example in Australia, a 'vertical scale' (where a movement along the scale between

[33] Ofsted, *Inspection 2012: proposals for inspection arrangements for maintained schools and academies from January 2012,* (2011).

any two equally spaced points must reflect similar levels of progress) is created by testing several year-groups, using some common questions to link scores on each test together. A particular question might be considered difficult for a Year 3 pupil, but much easier for a Year 5 pupil. Although this is technically defensible, it does require tests at more regular intervals than we currently have in England.

In England, we currently use National Curriculum levels as a scale against which to measure progress. However, as stated later in this chapter, concerns have been raised as to whether the levels, as they currently exist, are appropriate as a true vertical scale. **We recommend that, as part of the review of the National Curriculum, consideration is given to creating a more appropriate 'vertical scale' with which to measure progress.**

Enabling benchmarking of schools

One of the key uses we have identified for statutory assessment data is enabling benchmarking between schools. We recognise that effective benchmarking by school managers to evaluate performance and drive improvement will always be a complex process. A greater emphasis on progress and the introduction of broader accountability will enable better and fairer comparisons between schools.

Publishing a broader range of data will make it easier to compare schools with similar circumstances and challenges. However, we believe it is also important that every school should also be compared against the local and national average.

We believe that facilitating effective benchmarking by school managers requires additional tools and analysis. We feel that Raiseonline is an invaluable resource for school managers because of the detailed information it provides.

We welcome the Government's commitment to publish 'families of schools' data[34], which group schools into 'families' of 10 to 15 schools with similar intakes on the basis of prior attainment and socio-economic factors. We believe that schools and Ofsted should look to use this tool as they see fit.

[34] Department for Education, *The Importance of Teaching: the Schools White Paper 2010,* (2010).

International comparison studies

We believe accountability to Government, parents and the public also includes providing a picture of how our education system compares with education systems internationally. It is important that we continue to compare ourselves with other countries, as the challenge facing our education system is not only to improve year-on-year, but also to keep pace with the best education systems in the world.

Given the wide variation in assessment systems throughout the world, we do not believe that a statutory assessment system which is unique to this country can provide effective comparisons with other countries. We therefore consider that participation in the specifically-designed international comparison studies will continue to provide the most effective method of comparing standards of education in England against international standards.

We therefore recommend that England continues to participate in the following main international comparisons studies:

- **PISA** – the Programme for International Student Assessment is conducted every 3 years, assessing performance across reading, mathematics and science with one of those areas being the major focus in each cycle. Participating pupils are aged 15.
- **PIRLS** – the Progress in International Reading Literacy Study is conducted every 5 years, measuring trends in literacy achievement and collects information on policy and practices. Participating pupils are aged 10.
- **TIMSS** – the Trends in International Mathematics and Science Study is conducted every 4 years, providing data about trends in achievement over time and collecting information on quantity, quality and content of instruction. Participating pupils are aged 10 and 14.
- **TALIS (from 2012)** – the Teaching and Learning International Survey has been designed to provide data and analyses on the conditions needed for effective teaching and learning in schools.

We understand the Government proposes to make it a statutory duty for maintained schools to participate in international comparisons studies such as PISA, TIMSS, PIRLS and TALIS if selected in the samples. This is to ensure that participation rates for the studies are always met and data is not invalidated – on a number of occasions, England has struggled to secure sufficient schools and pupils to take part in international comparison studies. **We welcome the proposal in the 2011 Education Bill to make participation in international comparison studies mandatory for those schools selected.**

Chapter 3 – Statutory Assessment

Much of the evidence received by the Review has discussed the way in which current test data is used (and particularly 'high stakes' accountability). However, we have also heard a great deal of feedback with concerns over the way in which this data is generated. This chapter outlines the context of statutory assessment in England and the main concerns about it, before discussing both summative teacher assessment and external testing. It sets out the role each should play in a system of statutory assessment designed to achieve the purposes we set out in the Introduction.

History of statutory assessment

Given that England's system of statutory assessment has evolved over the course of more than two decades, it may be helpful briefly to trace this history to understand how the current arrangements have been formed[35].

The 1988 Education Reform Act introduced the National Curriculum and required that there should be "*arrangements for assessing pupils at or near the end of each key stage for the purpose of ascertaining what they have achieved in relation to the attainment targets for that stage*"[36]. The initial design of these arrangements was developed by the Task Group on Assessment and Testing (TGAT), a panel of experts chaired by Professor Paul Black. They proposed a standard national system of assessment designed to be formative as well as summative, covering all subjects (not just the core) with a single scale (the 10 level scale) for measuring pupil progress[37]. TGAT envisaged a system of teacher assessment in most Attainment Targets within each subject, supported by one or more Standard Assessment Tasks (SATs) administered by teachers. Evidence from teacher assessment and SATs was to be aggregated; where they differed, SAT results were to be preferred.

The first statutory assessments under the new National Curriculum took place in 1991, at the end of Key Stage 1. The number of Attainment Targets (each underpinned by a series of statements) meant that the assessments needed to cover a very considerable amount of information for each individual pupil. This caused significant workload issues, requiring around 44 hours of teachers' time rather than the 30 set out in the original specification. In Chris Whetton's analysis, "*the SATs had attempted to meet some of the principles of sound classroom-based educational assessment by providing a means whereby all children could show their best, but this was unmanageable in a mass testing system. However, there were positive outcomes, which were largely associated with the effects on the curriculum and on teachers' knowledge and skills in assessment… Compulsory authentic assessment had an immediate effect. But for many, the gains in beneficial effects on the curriculum were outweighed by the manageability and reliability arguments*"[38].

The next statutory assessments to be developed and introduced were at the end of Key Stage 3. Teacher-marked tasks were replaced with tests, developed to a specification based on the curriculum programmes of study. Criticism of the statutory assessment system grew, prompting a boycott and a legal ruling that teachers should not be required to mark statutory assessments without additional pay. The

[35] A more detailed outline can be found in Whetton, C., 'A brief history of a testing time: national curriculum assessment in England 1989-2008', in *Educational Research 51.2*, 137-159 (2009).
[36] *Education Reform Act 1988* 1, 2 (2).
[37] Department for Education and Science, *National Curriculum Task Group on Assessment: A Report*, HMSO (1987).
[38] Whetton, C., 'A brief history of a testing time'.

consequent Dearing Review led to a number of changes to the underlying curriculum and, importantly, the introduction of external marking. In Whetton's view, this "*paradoxically took the system further from one of the original educational objectives of TGAT, that of the teachers' marking their own students' work and deriving useful feedback from the process*"[39].

The system of statutory assessment had therefore changed considerably by the time Key Stage 2 arrangements were developed. When the first statutory assessments at the end of Key Stage 2 were introduced in 1995, they followed the Dearing reforms and had been designed from the very start as written tests covering the reduced curriculum. The tests have essentially retained their original form to the present day. There is one important legacy of the original design of statutory assessment – tests are still frequently referred to as 'SATs' rather than 'National Curriculum Tests'.

Perhaps the most important change has been the increasing weight placed on the outcomes of National Curriculum Tests, particularly since the 1997 General Election. The Government made widespread use of a 'target' culture to drive school improvement, combined with significant support for schools through the National Strategies and a focus on addressing underperformance. The relentless demand to 'raise standards' (*recte* performance) each year, both for individual schools and across the education system, led to the statutory tests becoming 'high stakes' both for schools and for Government. As the weight placed on end of Key Stage testing increased, so did the pressure placed on them. Concerns over the way boundaries between level thresholds were set led to the 1999 Rose Inquiry[40] and a Statistics Commission report[41] into the appropriateness of using test results to measure national standards. The concerns this Review has heard are essentially those which have built up over the previous decade: that tests serve too many purposes, and that their use for accountability puts too much pressure on schools and pupils and leads to narrowing of the curriculum[42].

In 2005, in response to concerns that formal testing was not appropriate for very young children, tests at the end of Key Stage 1 were replaced with a system of teacher assessment informed by tests and tasks and subject to external moderation. While greater weight was given to teacher assessment, a NFER evaluation in 2006 suggested that, at least in the first year, many teachers were making limited use of their freedoms, and administering tests and tasks late in Year 2 to inform their summative judgements[43].

In the summer of 2008, significant delivery failures resulted in the late delivery of National Curriculum Test results to many schools and pupils, especially at Key Stage 3. This failure in delivery prompted considerable criticism of the testing process generally and even the principle of externally-marked tests. In response, the Government adopted the recommendations of the Sutherland Inquiry to safeguard test delivery[44]. It also discontinued statutory tests at the end of Key Stage 3, accepting that GCSE results were the natural focus for school accountability. Following the report of the Expert Group on Assessment[45], published in 2009, full cohort testing of science at the end of Key Stage 2 was also discontinued, leaving schools to report summative teacher assessment to parents, while administering National Curriculum Tests to a sample of 750 schools to monitor national standards.

---

[39] Whetton, C., 'A brief history of a testing time'.

[40] Rose, J., *Weighing the baby: The report of the independent scrutiny panel on the 1999 Key Stage 2 National Curriculum tests in English and mathematics*, London, DFEE (1999).

[41] Statistics Commission, *Measuring standards in English primary schools: Report no.23*, London (2005).

[42] Whetton, C., 'A brief history of a testing time'.

[43] Reed, M., and Lewis, K., *Key Stage 1 evaluation of new assessment arrangements*, QCA (2005).

[44] Sutherland, S., *The Sutherland Inquiry: An independent Inquiry into the delivery of National Curriculum tests in 2008: A report to Ofqual and the Secretary of State for Children, Schools and Families*, TSO (2008).

[45] DCSF, *Report of the Expert Group on Assessment*, (2009).

In support of a greater focus on personalised learning, in January 2007 the Government consulted on a 'Making Good Progress' pilot, designed to explore new ways "*to measure, assess, report and stimulate progress so that no child is left behind*"[46]. An important product was the development and piloting of Single Level Tests, designed to be administered over a Key Stage to confirm whether a pupil was working at the level indicated by teacher assessment[47]. This was followed by the 'Assessment for Learning' strategy, investing £150 million over 2008-2011, encouraging schools to work together to develop teachers' assessment skills. The Government also published *Assessing Pupils' Progress* (APP), a set of materials for schools to use optionally to support teacher assessment in reading, writing, speaking & listening, mathematics and science for Key Stages 1, 2 and 3, providing criteria against which judgements can be made about levels and sub-levels.

In spring 2010 NAHT and NUT balloted their members on action to boycott the delivery of end of Key Stage 2 National Curriculum Tests. They argued that it was not appropriate to use a single week of externally-marked tests to hold schools publicly accountable. As a result, 4,005 schools (26% of those expected to do so) did not administer the tests. The considerable strength of feeling demonstrated by this action contributed directly to this Review being established.

Perceptions of current statutory assessment

This Review has considered a good deal of evidence and feedback about the system of statutory assessment in England, which we have outlined in our *Progress Report*[48]. A great deal of attention has focused on the role of end of Key Stage 2 National Curriculum Tests in particular.

It is important to stress that, while many respondents have criticised the current system, this criticism is generally not directed at the concept of testing. Many respondents (including the main teaching unions) have explained that their concerns are not principally about National Curriculum Tests themselves, but about the way that test results are used in the wider accountability system. When the online call for evidence asked which parts of the current system most needed to change, the most popular answer (50% of respondents to the question) argued that league tables should be abolished or tests no longer used to compile them; only 33% suggested that end of Key Stage 2 tests should be abolished. One respondent to the online call for evidence summarised the problem as follows:

> "*The high stakes nature of how the results are used is the issue, not the tests themselves. This pressurises schools into narrowing the curriculum. As it is so easy to measure a school purely on objective numerical results it is used as the main measure of a school's success by Ofsted. Consequently many schools focus on preparing the children just to score well in the literacy and numeracy tests, not instilling a love of learning for its own sake or providing a broad and balanced curriculum*."

As a panel we recognise the strength of professionals' feeling about the future of statutory assessment. The 2010 boycott demonstrated how strongly many heads feel about statutory assessment data and how it is used in the accountability system. The evidence we have received suggests that the fundamental concerns are with the way in which school accountability data is used. We have therefore recommended a number of reforms which would significantly improve the way in which primary schools are held accountable, through which we believe we can address the imbalances and perverse incentives in the school accountability system.

---

[46] DfES Consultation, *Making Good Progress: How can we help every pupil to make good progress at school?*, (2006).
[47] The Single Level Test pilot is discussed in Chapter 4.
[48] Bew, P., *Review of Key Stage 2 testing, assessment and accountability: Progress Report*, (2011).

However, we recognise there are also significant concerns with the current statutory assessment system itself, which we now move on to address.

The purposes of statutory assessment

When asked what they considered to be the main purposes of externally-marked tests, 45% of respondents to the online call for evidence (the greatest proportion who answered that question) chose 'to help teachers to set expectations and inform them about the performance of pupils'. This is considerably higher than the 21% who selected 'for government to measure the national performance', or the 14% who selected 'to enable parents and the public to compare levels of achievement across schools'.

We agree entirely that helping teachers to set expectations and inform them about the performance of pupils in order to support teaching and learning is crucially important. We believe that this purpose is best served through ongoing non-statutory teacher assessment, which is integral to good teaching and learning. Good teaching is wholly dependent on good assessment at every point, including throughout every lesson. It is clear to us that this type of assessment will be by far the most appropriate means of supporting the learning of each individual pupil. Teachers are uniquely well placed to assess their pupils and identify how best to help them improve.

We do not wish this Review to prescribe how teachers should carry out their day to day assessment. However, we recognise that there are links between statutory summative assessment and ongoing formative assessment, and that some of our recommendations will depend on schools having formative assessment arrangements in place.

Instead we have focused on statutory assessment and system-wide recommendations. We have identified the following three key uses for the statutory assessment system in addition to the purpose defined in primary legislation ("*to ascertain what pupils have achieved in relation to the attainment targets for that stage*"[49]):

- holding schools accountable for the attainment and progress made by their pupils and groups of pupils;
- informing parents and secondary schools about the performance of individual pupils;
- enabling benchmarking between schools; as well as monitoring performance locally and nationally[50].

The uses of statutory assessment data set out above clearly differ from those highlighted by many respondents in their feedback to the Review. We would like to be clear that the difference in uses is no reflection of their relative importance, but instead a reflection of what we believe can best be achieved through statutory assessment, and what should be achieved by teachers assessing their pupils on an ongoing basis.

The three main uses we have chosen concentrate on accountability to parents, the public and local and central Government. An important consequence of focusing on accountability is that the results of statutory assessment need to be considerably more robust and reliable. Any system producing data used for school accountability has a much higher burden of proof than one used internally.

[49] *Education Act 2002*, section 76.
[50] These purposes are discussed further in Chapter 1.

Any assessment requires a degree of compromise between validity, reliability and manageability. Validity and reliability have specific technical meanings which we feel should be explained clearly[51]:

- **Validity** is concerned with *what* is assessed: the extent to which it measures what it is intended to measure and whether the evidence and theory support the intended interpretation of the outcomes.
- **Reliability** is concerned with *how well* an assessment measures whatever it is measuring: the degree of consistency of the measurement and the probability that there would be the same outcome if the assessment was repeated.
- **Manageability** describes the impact the assessment has on the pupils undertaking it and the assessors who oversee it.

The validity of an assessment depends on the confidence which can be placed on its underlying processes: its purposes (are they clear?); its construct (does it actually assess the curriculum content it intends to?); its administration (are conditions consistent? is marking consistent?); and the dependability of results (is there bias? are pupils misclassified?)[52]. In essence, validity means "*does the assessment allow you to draw the conclusions you intend to*?"[53]. It can be difficult to quantify validity, since it will depend on a number of judgements.

The reliability of an assessment "*refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated… It concerns the likelihood that students would have received different results if they happened to have been allocated a different test date, a different version of the test, a different marker and so on. Reliability is about quantifying the luck of the draw*"[54].

It is important to note that reliability is not the same as accuracy. Some factors (for example tasks which give particular groups of pupils an advantage or disadvantage) can introduce error into an assessment on a systematic basis[55]. Since these factors tend to have the same effects repeatedly, they may not affect reliability in its technical sense (meaning the consistency of outcomes). An assessment can therefore be very reliable without being accurate.

A system of formative assessment, used internally within schools, usually prioritises validity whilst trying to ensure within-school consistency (rather than formal notions of reliability). A system of summative assessment used to compare schools and hold them accountable should prioritise reliability, whilst maintaining validity, so that fair comparisons are made between schools in different areas. The Government must be able to demonstrate that any system of statutory assessment is sufficiently reliable for the purpose of providing data for school accountability.

In determining the reliability of assessments and ensuring appropriate inferences are made from the outcomes, some have argued that the use of confidence intervals might be helpful. Confidence intervals are a way of demonstrating the impact of measurement error on results and are used in the international comparison surveys to enable comparisons between participating countries. However, there are concerns that confidence intervals may be confusing for schools and parents. **We recommend that further work is carried out to determine whether the use of confidence intervals would promote greater understanding of the outcomes of statutory assessment.**

---

[51] The following discussion draws heavily on the papers by Stobart and Newton in *Educational Research* 51:2 (cited below).
[52] Based on Stobart, G., 'Determining validity in national curriculum assessments', in *Educational Research* 51:2, 161-79 (2009).
[53] Stobart, G., 'Determining validity'.
[54] Newton, P.E., 'The reliability of results from national curriculum testing in England', in *Educational Research* 51:2, 181-212 (2009).
[55] Newton, P.E., 'The reliability of results'.

Summative teacher assessment and school accountability

Much of the evidence and feedback we have heard calls for externally-marked tests to be wholly replaced with a system of summative teacher assessment. Proponents of teacher assessment argue that using it for school-level accountability (in place of the current tests) would demonstrate trust in teachers' professional skills, remove incentives for inappropriate test preparation or narrowing of the curriculum, and avoid judging schools through a 'snapshot' of data from one week of tests.

In a joint policy statement[56], NAHT, NUT and ATL argued that "*teacher assessment, when it is subject to proper external checks and is carried out by well-trained professionals, has clear advantages over national curriculum testing… it is a more valid form of assessment, because a much wider range of pupils' learning, over a much longer time period, can be evaluated by the teacher than is possible through a few short, one-off, tests*". They propose a process of moderation that is "*teacher-led and locally organised, but accompanied by a resource which supports national standardisation: a national bank of assessment materials from which teachers can choose to draw to check their assessments.*" So that the Government can monitor national standards over time, 'low stakes' sample testing could be introduced.

The principal arguments for such a change are that, however good National Curriculum Tests are, they reflect a lack of trust in teachers, lead to excessive and educationally damaging test preparation (which may be stressful for pupils), and mean that the entirety of a school's achievements are reduced to the performance of 11 year-olds during a single week in May. It is argued that a system using only teacher assessment could solve all three problems at a stroke. Advocates suggest that greater use of summative teacher assessment would provide a more rounded and balanced picture of each pupil's attainment, empower the profession and act as a spur to develop the profession's assessment skills.

We recognise that these proposals have received support from a substantial number of primary teachers and heads. 2,743 primary heads and teachers (83% of respondents) replied to the online call for evidence to say that schools' own teacher assessment should be used to compare schools and hold them accountable. The following quotations from the online call for evidence indicate their reasoning.

> "*Teacher assessment based on detailed knowledge of pupils, supported by benchmarked assessment materials is a far more accurate means of measuring progress and attainment. This also creates meaningful data for the parents of the child to hold the school directly to account.*"
>
> "*Data that schools are judged by should not be solely based on one test, taken during one week, leading to an extremely high stakes inspection and accountability regime. The system was appropriate in the 90s when it was first conceived, and standards have improved significantly as a result, but as a profession we are so much more skilled now in our teacher assessment, and the testing system is so clumsy*."

Given the support for these proposals, we have spent a considerable amount of time considering them, particularly in relation to the purposes of statutory assessment we have outlined. One question we have constantly referred to is whether a system entirely using summative teacher assessment is sufficiently reliable for the purpose of providing data for school accountability. The next section explores this in some detail.

[56] ATL, NAHT, NUT, *Common Ground on Assessment and Accountability in Primary School,* (2010).

Reliability of summative teacher assessment

A crucial paper by Professor Harlen reviewed 30 research studies into summative teacher assessment judgements[57]. She cautioned that:

> *"Teachers should be made aware of the sources of bias in their assessments, including the 'halo' effect, and school assessment procedures should include steps that guard against such unfairness*"[58].

A number of research studies indicate that teacher assessment tends to under-state the achievement of pupils from minority groups[59]. Recent research by Professor Burgess suggests that "*on average, Black Caribbean and Black African pupils are under-assessed relative to white pupils, and Indian, Chinese and mixed white-Asian pupils are over-assessed… When forming an assessment of a pupil's likely progress, teachers use information on the past performance of members of that group in that school from previous years*"[60]. This study warns that relying solely on teacher assessment would disproportionately affect ethnic minority pupils. It suggests that replacing external tests with a system wholly of summative teacher assessment would reduce the proportion of ethnic minority pupils attaining level 4 by between four and seven percentage points, compared to a three percentage-point reduction for white pupils[61]. This indicates that external testing in some way protects pupils from typically low-attaining groups from subconscious assumptions. Burgess and Greaves reach an unequivocal conclusion:

> "*It is argued that pupils are subjected to too many written tests, and that some should be replaced by teacher assessments… The results here suggest that might be severely detrimental to the recorded achievements of children from poor families, and for children from some ethnic minorities… Given that 'setting' in secondary school classes may depend on earlier recorded attainment and that motivation may also be affected by a lower level, the use of assessment rather than testing may increase attainment gaps between ethnic groups later in academic life*" [62].

As Harlen summarises, "*there is bias in teachers' assessment (TA) relating to student characteristics, including behaviour (for young children), gender and special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills*"[63]. This evidence suggests that certain

---

[57] Harlen, W., *A Systematic Review of Evidence of the Impact on Students, Teachers and the Curriculum of the Process of Using Assessment by Teachers for Summative Purposes in Research Evidence,* EPPI-Centre, Social Science Research Unit, Institute of Education (2004).

[58] Harlen, W., *A Systematic Review of Using Assessment by Teachers for Summative Purposes*. Harlen cites a number of studies, including Bennett, R.E., Gottesman, R.L., Rock, D.A., Cerullo, F., 'Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill', in *Journal of Educational Psychology* 85, 347-356 (1993); and Brown, C.R., 'An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England', in *Studies in Educational Evaluation* 24, 87-98 (1998); Brown, C.R., Moor, J.L., Silkstone, B.E., Botton, C., 'The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations', in *Assessment in Education* 3, 377-391 (1996).

[59] Strand, S., *Minority ethnic pupils in the Longitudinal Study of Young People in England (LSYPE),* DCSF (2007).

[60] Burgess, S. and Greaves, E., *Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities*, Centre for Market and Public Organisation, University of Bristol (2009).

[61] Burgess and Greaves, *Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities.*

[62] Burgess and Greaves, *Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities.*

[63] Harlen, W., *A Systematic Review of Using Assessment by Teachers for Summative Purposes*.

pupils could be disadvantaged by a system which relied more heavily on teacher assessment.

There is also a risk that teachers would come under increasing pressure to make generous assessments for some or all of their pupils. Those on the borderline between levels (especially levels 3 and 4) might be given additional teaching at the expense of other pupils, producing extra work to justify a higher result – which would fail to remove an important perverse incentive in the current system.

There are therefore clear risks that summative teacher assessment will not be sufficiently reliable in a technical sense – i.e. that judgements will not be made consistently by teachers across the country. Moreover, the research indicates that pupils who are most likely to under-perform would be most vulnerable to under-assessment.

We are particularly aware of this risk given that Ofsted's 2010 HMCI report[64] shows that "*assessment continues to be an area in which schools need to improve*". Only 53% of schools are judged to be 'good' or 'outstanding' at assessment. When combined with the risk of subconscious bias, we would need considerable caution before moving to a complete reliance on summative teacher assessment for accountability purposes.

In his discussion of validity of National Curriculum assessments, Stobart concluded that "*test results are used because they represent an external measure of pupil attainment. In this high stakes context teachers' assessments would be suspect given the importance of good results to a school*"[65]. In his view, externally-set and marked tests represent the most valid way of assessing school performance: "*in a culture of simplistic performance measures and targets it is not easy to envisage alternative approaches that would satisfy these demands*" [66].

Moderation and summative teacher assessment

The joint NAHT, NUT and ATL statement recognises that "*teachers have a tendency to stereotype pupils according to ethnicity, but this bias can be overcome by two measures which will also have other important system benefits: moderation and professional development*"[67]. They propose a process of moderation that is "*teacher-led and locally organised, but accompanied by a resource which supports national standardisation: a national bank of assessment materials from which teachers can choose to draw to check their assessments.*"

We understand the huge value of moderation in terms of teacher CPD and providing opportunities for teachers to develop their assessment skills by learning from each other, which we cover at the start of the next chapter. However, this Review has heard evidence from a number of assessment experts about the limits of moderation for making teacher assessment judgements more robust and reliable.

Professor Wiliam gave his view that the experience of the last ten years does not suggest that moderation would make summative teacher assessment sufficiently reliable to be used for school-level accountability. Professor Tymms felt that moderation of the Early Years Foundation Stage Profile (EYFSP) and Key Stage 1 does little to alter teachers' judgments. Lord Sutherland, former chair of the Chartered Institute of Educational Assessors (CIEA), was concerned that the costs of strengthening teacher assessment to such an extent that it could replace external assessment would be very significant and probably prohibitive. A seminar attended by a wide range of assessment experts organised by NFER in 2009 concluded that

[64] Ofsted, *The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2009/10*, HMSO (2010).
[65] Stobart, G., 'Determining validity'.
[66] Stobart, G., 'Determining validity'.
[67] ATL, NAHT, NUT, *Common Ground on Assessment and Accountability*.

"*if the purpose* [of summative teacher assessment] *is about making judgements on teachers or institutions, there needs to be such stringent moderation practices that the assessment is no longer teacher assessment or it becomes unaffordable*" [68].

There are additional concerns about moderation of teacher assessment judgements in terms of teacher workload. External moderation arrangements will require teachers to have evidence for the judgements they have made, which may impose significant additional burdens on teachers, in particular through pressure for greater record-keeping. A key concern raised through Dame Clare Tickell's Review of the Early Years was that practitioners felt obliged to keep evidence to support all of their decisions to satisfy local authority moderators and Ofsted inspectors[69]. NASUWT has echoed this concern, suggesting that their Welsh members have seen an increase in workload since the discontinuation of external testing. We are also mindful that the current system of externally-marked testing was introduced to replace a system of teacher assessment which had proved itself to be exceptionally burdensome.

The evidence presented here does not suggest to us that moderation would address the considerable risks around reliability of moving to a system based entirely on teacher assessment.

It has also been suggested that teachers could use a range of tests to confirm summative teacher assessment judgements, thereby making them sufficiently reliable to be used for school accountability. Proposed models include using whole-cohort tests to moderate summative teacher assessment judgements for each pupil or having a sample of pupils sit tests to gauge the overall accuracy of teacher assessment. Test-based moderation assumes that pupils will perform similarly in tests (which capture a snapshot of attainment) and in teacher assessment (which draws on a much broader base of evidence). This is a fundamentally flawed assumption. While there is frequently a close correlation between teacher assessment and test results, it does not follow that a discrepancy means one or other is 'wrong'.

Because tests and teacher assessment draw on different evidence bases, it is perfectly legitimate for their outcomes to be different. For example, a pupil could consistently demonstrate strong level 4 attainment over Year 6, but perform particularly well on the test and be awarded level 5. Harlen's study reached a clear conclusion that "*the essential and important differences between TA and tests should be recognised by ceasing to judge TA in terms of how well it agrees with test scores… Teachers should not judge the accuracy of their assessments by how far they correspond with test results, but by how far they reflect the learning goals*"[70].

We do not believe that tests should be used to validate summative teacher assessment to provide data for school accountability. It is for this reason that we do not believe Key Stage 2 statutory assessment should follow the same model as Key Stage 1 statutory assessment. While we believe it is appropriate at Key Stage 1 for teacher assessment to be informed by tests, we do not feel this model would be appropriate for the purposes we have outlined for Key Stage 2 statutory assessment.

Summative teacher assessment in Wales

In 2005 the Welsh Assembly Government replaced statutory testing with a system of teacher assessment (moderated across local clusters of schools) designed to focus on individual pupils and smooth the transition from primary to secondary education. It

[68] Parkes, C., and Maughan, S., *Policy and research seminar on Methods for Ensuring Reliability of Teacher Assessments: Proceedings*, NFER (2009).
[69] Tickell, C., *The Early Years: Foundations for life, health and learning, An Independent Report on the Early Years Foundation Stage to Her Majesty's Government*, (2011).
[70] Harlen, W., *A Systematic Review of Using Assessment by Teachers for Summative Purposes*.

offers a valuable case-study of the impact of such a change, which we have considered with great interest.

Since 2005 there have been no statutory tests or tasks in Welsh schools at Key Stages 2 or 3. Instead, schools are required to make summative teacher assessment judgements for all eligible pupils at the end of Key Stage 2. There is also standardisation and moderation of examples of pupils' work, both within schools and by clusters of schools (including secondary schools). National exemplification and guidance on moderation has been made available since 2008.

Estyn (the Welsh inspectorate) recently published a survey[71] into the effectiveness of teacher assessment. They found that "*overall, primary and secondary school teachers are becoming more confident about their understanding of the characteristics of pupils' work that demonstrate the NC level descriptions. Nevertheless, different perceptions as to what constitute appropriate standards in the core subjects at different levels in end-of-key stage 2 and key stage 3 assessments continue to exist.*" However, "*Estyn's evidence from school inspections across Wales consistently indicates that about a quarter of schools inspected each year have shortcomings in aspects of assessment… Estyn's inspections of schools consistently indicates that assessment is one of the weakest areas of work in schools. There are also weaknesses in the use that secondary teachers make of assessment information from primary schools.*"

Estyn concludes that "*in KS2, confidence in the system of assessment, and in particular in the consistency of its application across different school cluster groups, is not as reliable as it should be because there is not, as yet, an effective process in place to moderate these judgements externally. In addition, as it currently operates, the process does not actually verify that the levels finally awarded to examples of pupils' work at the end of key stage 2 are accurate across the full range of individual schools, clusters of schools and local authorities. This means that a level 2, or any given level, will not mean the same in different parts of the country.*" The absence of external moderation of Key Stage 1 teacher assessments in Wales "*makes the reliability of value added and benchmark measures between the key stages uncertain.*"

Professor Daugherty (chair of the Daugherty Assessment Review Group[72] which reported in 2004) observed that the reforms introduced in Wales had smoothing of the transition between Years 5 and 7 ('crossing the divide') as their primary purpose[73]. He made it clear that the current system of national assessment in Wales, focusing on the individual pupil and based on summative teacher judgements moderated across local clusters of schools, cannot and should not be expected to supply aggregate data that is robust enough for accountability purposes. Professor Daugherty has suggested separately that it is not appropriate to use teacher assessment for school based accountability[74].

In a speech on 2nd February 2011, Leighton Andrews (Minister for Education in Wales) acknowledged systematic failings in Welsh education, and made a commitment to introduce "*a national system for the grading of schools to be operated by all local authorities… all schools will produce an annual profile containing*

[71] Estyn, *Evaluation of the arrangements to assure the consistency of teacher assessment in the core subjects at key stage 2 and key stage 3,* (2010).
[72] Daugherty, R., *Learning Pathways through statutory assessment: Key Stages 2 and 3, Daugherty Assessment Review Group Final Report*, (2004).
[73] Summarised in Daugherty, R., 'National Curriculum Assessment in Wales: adaptations and divergence', in *Educational Research* 51(2), 247-250 (2009).
[74] Daugherty, R., 'Designing systems of teacher based summative assessments', in Parkes, C., and Maughan, S., *Policy and research seminar on Methods for Ensuring Reliability of Teacher Assessments: Proceedings*, NFER (2009).

performance information to a common format" and to introduce "*a national reading test which will be consistent across Wales*" [75].

We believe the evidence from Wales and the current shift away from a reliance on teacher assessment in Wales suggests external testing should play an important role in a statutory assessment system which aims to provide data for accountability purposes.

Summative teacher assessment – conclusions

We have considered the evidence for using summative teacher assessment as the basis for school accountability in some detail. The strength of support for it from teachers and heads is very significant, with over 80% of respondents to the online call for evidence arguing that it should replace externally-marked testing.

However, we recognise the important concerns that judgements may not be reached through consistent use of national standards (affecting validity) and that judgements may be affected by extraneous factors such as pupil behaviour (the risk of subconscious bias, affecting reliability). The evidence that subconscious bias disproportionately affects some of the most vulnerable pupils is especially worrying. While we acknowledge that moderation has an important place in improving the quality of teacher assessment, the evidence we have considered does not suggest it would make summative teacher assessment reliable enough to become the only source of data for school accountability.

We feel that a system entirely based on teacher summative assessment would not be sufficiently reliable for the purpose of providing school accountability data. **While we believe ongoing and high quality assessment is crucial to ensuring pupils make good progress, we do not believe that schools should be held accountable through a system wholly based on moderated teacher assessment**.

[75] Leighton Andrews, 'Teaching makes a difference', Reardon Smith Theatre, Cardiff, 2 February 2011.

Key Stage 2 National Curriculum Tests

The evidence received by the Review concerning the National Curriculum Tests themselves has largely been positive.

In discussion, NAHT noted that tests are most effective where there are 'right' and 'wrong' answers, and consequently mathematics and some elements of science lend themselves to written tests in a way that writing composition does not. The following quotations from the online call for evidence give a sense of respondents' views on the current end of Key Stage 2 tests.

> "*The production of standardised tests which are objectively marked by external assessors is good. It allows complete confidence and some measure of objectivity in determining assessments for individual pupils*."

> "*A well designed test is a very efficient way of measuring attainment. APP is far too bureaucratic to be considered an alternative. The profession is familiar with testing over decades. Keeping tests would therefore be good but only if the results are not used to compile school by school league tables*."

> "*The tests themselves could be useful for school's own use and for any government to be able to gain a snapshot of attainment at end of KS2. It is the current use (or abuse) of the test results that most schools have a problem with, because there are potentially significant consequences for headteachers and schools with 'poor' results*."

NFER, who have been contracted to develop the English and science sample tests, argued that "*the current tests could be considered the best tests of their type and are developed using the most thorough test development approaches available.*"

Nonetheless, there are questions around the validity and reliability of the Key Stage 2 tests. Any assessment requires a degree of compromise between validity, reliability and manageability. However, some respondents to the Review believe that the current end of Key Stage 2 National Curriculum Tests do not strike the right balance.

Reliability and validity of National Curriculum Tests

The reliability of National Curriculum Test results has been questioned, with debate centring on the quality of marking, even though this is only one facet of technical reliability. It may be worth highlighting at this stage that terms like 'reliability', 'error' and 'misclassification' have specific technical meanings in an assessment context. Ofqual's report on reliability of National Curriculum Tests noted that "*the word 'error' in its daily use is constantly associated with human mistakes, while its technical meaning in educational measurement implies deviation of scores on a test from some notional number when the measurement procedure is repeated*"[76]. It is easy for technical discussions of 'reliability', 'measurement error' and 'misclassification' to be misinterpreted.

This is perhaps most apparent in the case of research into misclassification by Professor Wiliam in 2000 which suggested that 32% of pupils could be given the wrong National Curriculum level, a figure which has frequently been cited[77]. However, analysis by Dr. Newton indicates that *"the estimates of 'incorrectness' derived from entering empirical data into Wiliam's simulation seem to be almost twice as high as those derived directly from the empirical data*"[78]. Ofqual's recent report[79]

[76] Opposs, D. and He, Q., *The Reliability Programme: Final Report*, Ofqual (2011).

[77] Wiliam, D., 'Reliability, validity, and all that jazz', *Education 3-13,* 29(3), 9-13 (2000); Newton, P.E., 'The reliability of results' discusses the way this research has been used and its implications.

[78] Newton, P.E., 'The reliability of results'.

indicates that the 2010 National Curriculum Tests had a misclassification rate of 15% in English, 13% for the science sample test and 10% in mathematics. Newton's research supports the conclusion that reliability is highest in mathematics, followed by science, reading and finally writing[80].

It is generally accepted that any test or examination, however well constructed, will always include a degree of measurement error. Therefore the margin of error for both pupils and schools needs to be considered. Professor Wiliam made the important point to the panel that misclassification can be both upwards and downwards, so in a large enough cohort any errors are likely to balance out and not have a statistically significant impact. Professor Black and his colleagues took a similar view when giving evidence to the Education Select Committee, suggesting that the inevitable concerns around reliability were not in themselves an argument against the use of formal tests, but rather a warning that they should be used with an understanding of their limitations[81].

We understand that, as with all tests where pupils are categorised, the level thresholds in Key Stage 2 tests mean that one mark can make the difference between one level and the next. That mark could be lost or gained through a pupil mis-reading an instruction in the test or making a fortunate choice in a multiple-choice question, or through slight variations in marking practice. These differences will be highly significant for the individual pupil (though they might be qualified by an accompanying teacher assessment judgement) but, at school level, they are likely to cancel each other out across the full cohort of pupils[82].

We believe that, for the purposes we have outlined which concentrate on school accountability, test outcomes represent a sufficiently reliable indicator of the overall performance of pupils at a school, particularly if they are looked at over a three-year period.

We recognise that there is a concern that the need for high reliability in National Curriculum Test results means that validity is compromised[83]. In practice, what is tested is limited to what can be most consistently marked. This means that important areas of the curriculum may be marginalised or excluded from end of Key Stage 2 testing – in particular use and application of mathematics, speaking and listening and the more practical aspects of science. If these topics do not appear in the 'high stakes' testing regime, schools may choose to give them low priority in their teaching. We therefore need to ensure that any future tests strike the right balance between reliability and validity.

Using National Curriculum Tests to measure standards over time

The third main use of statutory assessment data that we have outlined ("enabling benchmarking between schools, as well as monitoring performance locally and nationally") requires both comparing schools at the same point in time and monitoring trends over time. This presumes that statutory assessment in one year will produce results which are comparable to statutory assessment from another year.

In particular, Professor Tymms has questioned the appropriateness of using National Curriculum Test results to monitor national standards over time, suggesting that the

---

[79] Opposs and He, *The Reliability Programme: Final Report*.
[80] Newton, P.E., 'The reliability of results'.
[81] Black *et al* evidence to House of Commons Children, Schools and Families Select Committee, *Testing and Assessment: Third Report of Session 2007-08*, TSO, HC-169-II (2008).
[82] Newton, P.E., 'The reliability of results' discusses this issue, and the potentially different ways in which assessment experts and statisticians understand 'reliability'.
[83] Harlen, W., *The Quality of Learning: assessment alternatives for primary education*, Cambridge Primary Review Research Summary 3/4 (2007).

test results have over-stated increases in national pupil performance (for example in the period 1995-2000)[84].

The standards in National Curriculum Tests (i.e. the level of performance expected to achieve a particular level) are maintained over time through the level-setting process. This process (outlined in Ofqual's regulatory report[85]) employs a variety of statistical and judgemental techniques, including statistical equating, comparison of completed papers with those considered to be at the same level in previous years, and comparison of the national pupil cohort's performance with that in previous years. While this is a complex process, it appears to be effective: Ofqual's report on the 2010 National Curriculum Tests showed that the level-setting process worked well, concluding "*there can be confidence that standards were maintained*"[86]. The Statistics Commission's 2005 review of these issues (in response to concerns raised by Professor Tymms) found that *"test scores may not be an ideal measure of standards over time, but it does not follow that they are a completely unsuitable measure*"[87].

A number of respondents to the Review have suggested that a system of purposely-designed sample testing would provide a more efficient way of monitoring national standards. They argue that a 'low-stakes' sample test could be administered securely, allowing the same questions to be re-used year after year, providing a very robust measure of standards over time.

We believe that it is legitimate to use statutory assessment data, including from National Curriculum Tests, to make comparisons over time. We recognise the benefits of a system of national sample testing. However, we do not feel it would be cost-effective to introduce a system of sample testing in addition to whole-cohort National Curriculum testing.

External marking

An essential characteristic of end of Key Stage 2 National Curriculum Tests is the fact that they are externally marked – in contrast with the tests and tasks used in Key Stage 1 statutory assessment. In general, the evidence we have heard about tests specifically suggests that external marking is welcomed, as this response to the online call for evidence indicates:

> "*External marking is generally good as it creates impartiality and reduces pressure on staff*."

However, we have also heard views that external marking is unnecessary:

> "*External testing arrangements… are extremely cumbersome, heavily bureaucratic and unnecessarily expensive. Tests could be administered in schools and marked as Key Stage 1. External moderation could take place*."

25% of respondents to the online call for evidence were concerned that the quality of marking was either poor or at best variable. However, criticisms of external marking mostly focus on writing; for example, respondents to the online call for evidence commented:

---

[84] Tymms, P., 'Are standards rising in English primary schools?', in *British Educational Research Journal* 30, 477-494 (2004); Tymms, P., and Merrell C., *Standards and quality in English primary schools over time: the national evidence*, Primary Review Research Survey 4/1, Cambridge (2007).
[85] Ofqual, *National Curriculum Assessments Review Report: 2010 Key Stage 2 tests,* (2011).
[86] Ofqual, *2010 Key Stage 2 tests*.
[87] Statistics Commission, *Measuring standards in English primary schools: Report no.23*. London (2005).

> "*There are issues with the reliability and consistency of externally marked tests – particularly in relation to the writing assessment.*"
>
> "*The marking of writing is variable and relies too heavily on the interpretation and administration of a set of rules that in turn lead to distorted and contrived pieces of writing*."

External marking was introduced at Key Stage 3 following the 1994 Dearing Review and the 1993 High Court case where NASUWT successfully argued that teachers should not be expected to mark statutory tests without additional pay[88]. We are concerned at the potential workload implications of expecting teachers to mark statutory tests within a relatively short period of time at the end of Key Stage 2.

More importantly, external marking plays a crucial role in ensuring reliability of results, giving confidence that mark schemes are applied fairly and consistently. Markers receive training in the use of mark schemes and the quality of their marking is closely scrutinised. The stringent marking process has been explained by QCDA[89] and is described in Ofqual's annual reports on National Curriculum Testing. Ofqual's report on the 2010 tests concluded that "*observation of the marking process, through which the standards are applied accurately and consistently to each pupil's test scripts, demonstrated that compliance with the requirements… was high*"[90]. Removing external marking from tests used to hold schools accountable would therefore expose a 'high stakes' system to concerns of inconsistency and poor reliability. Given that the test results are used to compare schools' performance and build a picture of the national pupil cohort, it is essential that they are consistent and reliable. **We believe that a system of external marking is essential to ensure tests can provide data which is reliable and robust enough to be used for school accountability.**

[88] Dearing, R., *The National Curriculum and its assessment: Final report*, London (1994).
[89] QCDA, *Test development: Level setting and maintaining standards*, (2010).
[90] Ofqual, *2010 Key Stage 2 tests.*

Recommended reforms to statutory assessment

Both summative teacher assessment and external testing are important forms of assessment and both have a valuable role to play. Both have their strengths and both have their limitations. We do not believe that statutory assessment needs to rely only on one or other of these forms. We feel that decisions on whether a particular aspect of a subject should be tested should be determined by the educational merits of the specific case.

We believe that summative teacher assessment and externally-marked testing should not be seen in opposition to each other or used to judge each other. We believe that, since both draw upon different evidence bases, test and teacher assessment results can legitimately be different. Drawing on both can provide a combination of independent, external data and information which reflects a pupil's work across a year from teachers who know each child as an individual and have developed an excellent understanding of what they can do.

In reaching the following recommendations, we have considered the issues with the current end of Key Stage 2 National Curriculum Tests identified through our call for evidence. We have considered whether it is legitimate to use externally-marked tests to assess each core subject and what impact this may have on teaching. Where specific concerns have been raised, we have put forward proposals to address them. Our recommendations are guided by these factors, so that the form of assessment is determined by what we believe is most educationally appropriate.

It may be worth noting at this stage that we are aware some of our recommendations cannot be implemented straightaway due to the long lead-in times associated with the test development process. We are mindful that the development of high quality National Curriculum Tests requires a stringent test development process. We therefore recognise that the test development timescale places unavoidable practical constraints on the speed with which changes can be made to National Curriculum Tests.

Reading

The current reading tests have been reasonably well supported by respondents, although some specific concerns have been raised. Some respondents felt that the design of the current tests, while well suited to pupils working at levels 4 and 5 (around 85% of pupils), do not allow weaker readers to demonstrate fully what they can do. 27% of respondents to the online call for evidence recommended including a wider choice of topics, as pupils' interests could influence their comprehension. 22% felt that more time should be allowed for the tests, arguing that the test disadvantages children who are slower at processing written information, and therefore often do not finish the tests in the time allowed. These responses to the online call for evidence identify some of the issues:

> "*These tests are fine for able or more able children. Any child that has a slight difficulty with reading is automatically disadvantaged by the test procedure. They are unable to demonstrate their true knowledge and understanding that would be evident in a less stressful setup.*"

> "*Often the content is not relevant to many children because it is about topics which they know nothing about. Children who have wider life experiences and parents who encourage them to learn about the world around them can draw on these to support their answers. It puts many children at a disadvantage*."

Several suggestions were made for resolving these issues, including producing separate tiered papers covering levels 3-4 and 4-5; organising the questions to get steadily more difficult; and increasing or removing the time limit. One respondent wished to "*split tests into reading ability (mechanical skills of reading) and reading*

for understanding (especially for children working at L3)." Others criticised the amount of writing: "*a child can be a level 4 reader but may not have as fast a writing speed which will hinder completion.*"

The feedback we have received suggests that, while specific issues must be addressed, there are no fundamental concerns with an externally-marked test of reading. **We believe it is legitimate to use an externally-marked test to establish how well a pupil can read and comprehend a passage of text within a finite period of time**. If pupils are to access the secondary curriculum, it is essential that they are confident and fluent readers. **We recommend that reading should continue to be subject to externally-marked testing**.

We recognise the concern that the current reading tests may not allow lower-attaining pupils to demonstrate fully what they can do. We believe that the reading test should be accessible to all pupils. It may be possible to achieve this by adjusting the balance of text and reading time, putting texts and questions in a clear order of difficulty, and ensuring that the texts themselves are accessible to all pupils. **We recommend that, as new reading tests are developed, these suggestions should be incorporated in the new test design**.

In addition, we feel that the reading test should be, as far as possible, a test of reading rather than writing. At present many questions can be answered by marking the relevant choice, but some require longer answers to demonstrate more advanced comprehension of the text. **We recommend that the number of written responses in the reading test should be kept under review so that the test is, as far as possible, a test of reading**. We do acknowledge that some written responses may be needed, particularly in more demanding questions.

We believe it is most important for every pupil to leave Key Stage 2 as a fluent and confident reader, ready for secondary education. We believe that the most crucial aspects of reading at the end of Key Stage 2 are accuracy (decoding familiar and unfamiliar words correctly), fluency (speed and confidence) and comprehension (drawing meaning from text). Therefore the end of Key Stage 2 reading test should demonstrate each pupil's accuracy, fluency and comprehension. **We recommend that the Government should consider the skills which should be assessed by the reading test and we recommend that these skills should be brought out more clearly in the design of future tests that assess the new National Curriculum**.

We also suggest that the current assessment foci in reading could be reconsidered – they currently encourage schools to concentrate teaching on interpreting texts and understanding authorial intent. We feel that there is a possibility that this may lead to unhelpful test preparation. We feel there is a risk that being forced to over-interpret texts may take pupils away from reading for pleasure and could potentially restrict their love of reading. Pupils at Key Stage 2 should concentrate on reading fluently and regularly; and we believe it is essential that they enjoy their reading and read widely and often with texts becoming increasingly challenging.

Writing

We have heard a great deal of criticism of the current writing National Curriculum Test. 43% of respondents to the online call for evidence felt the writing National Curriculum Tests are 'inadequate', and a further 33% felt they are 'not very effective'. The following quotes reflect the views of many respondents

> "*This is the poorest of the tests... High quality writing is often marked down as certain aspects that are expected and used for scoring may not appear, yet the content can be high level. The genre highly influences the writing outcome as some seem to draw out better writing than others which can cause fluctuations in schools' results*."

"*The marking is subject to the whims of individual examiners. It is a subjective exercise and therefore pointless. It also currently is skewed towards creativity rather than writing skills such as punctuation etc… It rewards prescriptive teaching*."

"*The expectation on some children to write their best piece of work with little preparation in a stressful situation is unrealistic, then when marked, results are clearly inaccurate (sometimes higher/sometimes lower*)."

"*Writing is the hardest thing to mark and must be very difficult to design each year... You could just test pupils' spelling, grammar and handwriting. This can be done using ICT. Creative writing is very difficult and expensive to judge*."

There are clearly significant issues with the current writing tests. Respondents feel they do not reflect classroom practice, whereby children take time with their writing and put effort into spelling, punctuation, grammar, vocabulary and handwriting. Others observed that many children produce their best work as part of a structured lesson following an inspired discussion or school trip, while it can be difficult to write creatively under pressured test conditions.

The unpredictability of the writing genres is a point of particular contention, especially where pupils respond in the wrong genre or it is felt that the test includes a 'more difficult' genre. Perhaps the most significant point is the frequently-made criticism over the inconsistency and subjectivity of the external marking. This has fundamental consequences for professionals' confidence in the writing tests, as one respondent observed: "*results are clearly inaccurate... It makes a mockery of pupil achievement measures.*"

We recognise that there are some elements of writing – spelling, grammar, punctuation, vocabulary – where there are clear 'right' and 'wrong' answers, which lend themselves to externally-marked testing. A spelling test currently forms 14% of the writing test. Internationally a number of jurisdictions conduct externally-marked tests of spelling, punctuation and grammar (sometimes termed 'English language arts'). These are essential skills and **we recommend that externally-marked tests of spelling, punctuation, grammar and vocabulary should be developed**. We suggest it may be appropriate for handwriting to be assessed in this externally-marked test too.

However, there is much more to 'writing' than spelling, punctuation, grammar, vocabulary and handwriting and we believe writing composition in a wide range of forms for different purposes and audiences should also be a core part of the statutory assessment of writing.

We believe that there is fundamental challenge with the marking of writing composition (extended writing of prose, verse, formal letters etc) because it requires a professional's judgement rather than being empirically 'right' or 'wrong'. Ofqual's regulatory reports have shown that the writing test is the least reliable National Curriculum Test[91]. We feel that marking of writing could be improved through the introduction of double marking and note that, internationally, England is unusual in having a single external marker for writing tests. For example, Massachusetts, USA has two independent markers for each writing paper, while Alberta, Canada encourages teachers to mark their schools' tests and make comparisons with the formal marking process. QCDA developed a technique of 'adaptive comparative judgement' as part of its Single Level Test pilot. This process repeatedly compares pairs of answers, with judges making holistic assessments about which is better to establish their relative quality; it achieved very high rates of technical reliability. However, while employing two or more markers may help improve the reliability and accuracy of marking, we feel that the criticism of the marking of writing is not principally caused by any faults in the current process, but is due to inevitable

[91] Ofqual, *National Curriculum Assessments Review Report: 2010 Key Stage 2 tests*, (2011)

variations of interpreting the stated criteria of the mark scheme when judging a piece of writing composition.

In addition to concerns about marking, the limited choice of genre in the current writing test is a cause of considerable concern to schools, since teachers feel that some genres are 'easier' or 'harder' for certain pupils, thus affecting their performance in the test. Given the 'high stakes' nature of the test, the choice of genres for the two writing tests in any particular year may make a great difference to a school's results, making for greater unpredictability of outcome. Experience of previous tests suggests that allowing pupils a choice of genre is unlikely to have a beneficial effect. It might encourage pupils to learn pre-prepared examples (both reducing the fairness of the test and rewarding an unhelpful form of test preparation). Professor Wiliam cautioned that lower attaining pupils tend not to choose genres which will allow them to show their full potential. We therefore do not believe this is the most effective solution.

We believe that it can be legitimate to assess writing composition through an externally-marked test. However, we share many of the significant concerns that have been expressed about the inherent challenge of marking writing tests, the impact of the choice of genres, and the feeling that, in comparison with other subjects that are tested externally, it is less valid to measure pupils' attainment on the basis of one test paper in May. **We recommend that writing composition should be subject to summative teacher assessment only**. This will encourage a broad range of writing over the course of Year 6, while avoiding the perverse incentives of the current system. It would allow Year 6 pupils to demonstrate what they can do across a range of genres, and would remove the inevitable disagreements about the marking of individual pieces of writing.

We believe this shift in the assessment of writing composition will help develop the creativity of the teaching profession. We want pupils to be taught a wide range of writing genres and to be encouraged to produce their best work each time they write rather than having strict time constraints. This is more likely to lead to a 'can do' attitude towards writing and greater enjoyment than is the case if teaching across the year is based on a build up towards the current test.

We know that teachers and head teachers throughout the country are capable of using a range of assessment techniques to assess each pupil's writing composition, and we want to give teachers the space to do so. We believe that, with this new approach teachers will be encouraged to approach writing composition in a richer and broader way, and will be more likely to differentiate teaching and learning depending on what is right for their pupils.

We are very conscious of the need for teacher assessment to be reliable and command public confidence. **We recommend that teacher assessment in writing composition should be subject to external moderation**. **We recommend that, if the moderator has concerns over the accuracy or reliability of the sampled teacher assessment judgements, they should be able to scrutinise additional evidence and, if they consider it appropriate, require the school to change the reported levels**.

We do not wish the moderation process to impose additional burdens on teachers. We do not feel it would be helpful or appropriate for schools to create portfolios of work or specially prepared 'show-piece' examples. We believe it would be most appropriate for moderation to review exercise books and other examples of marked written work for a range of purposes taken from the whole teaching year. This could include written work from other subjects as well as 'English'. If moderators review marked work produced in the course of everyday teaching in Year 6 (together with the teacher's associated comments) they will get a strong sense both of the pupil's attainment and the teacher's assessment skill, without creating any additional workload. We feel it may also be helpful for moderators to have the option of meeting the pupils whose work they have reviewed.

We accept that this form of moderation may not realise the professional development benefits which 'cluster' moderation can bring. Cluster moderation focusing on teachers learning from each other to develop their assessment skills is covered in the next chapter. We feel that it can play an important role in building a shared understanding of National Curriculum levels, but it would not be sufficiently robust to provide the only form of moderation.

We therefore recommend that writing should be assessed through a mixture of testing and summative teacher assessment. Due to its importance, **we believe that writing composition should always form the greater part of overall writing statutory assessment.** We recognise that we are recommending a very significant change to the statutory assessment of writing, addressing the profession's strongly-held concerns.

Speaking and listening

We recognise that speaking and listening are critical to the teaching and learning process. Children must be able to articulate their ideas and understand what they have been taught if teachers are to assess what they know.

Relatively little evidence has been presented to us concerning speaking and listening. While 90% of online call for evidence respondents agreed that there were elements of English which should only be teacher assessed, just 31% singled out speaking and listening:

> "*Speaking and listening would be best assessed by a familiar adult*."
>
> "*Speaking and listening can only ever really be teacher-assessed. However, when all the pressure is on the reading and writing tests, the danger is that S&L becomes a poor relative and is neglected*."
>
> "*Speaking and Listening should be dropped altogether from assessment. It has never been properly assessed anyway*."
>
> "*There could be an externally monitored speaking and listening test that is taped for children who have writing difficulties e.g. motor skills, dyslexia*."

We have heard no evidence to challenge the current arrangements of summative teacher assessment in speaking and listening. We acknowledge that its assessment has a relatively low profile in many schools when compared to reading and writing. **We recommend that teacher assessment of speaking and listening should continue**. It should be reported to parents and secondary schools and should continue to inform schools' overall teacher assessment of English. In view of the nature of speaking and listening, we do not feel that external moderation arrangements would be appropriate or proportionate; therefore, while speaking and listening should contribute to an overall teacher assessment of English, it may not be sufficiently reliable to be used as a measure of school accountability.

We recognise the importance of speaking and listening and the need for all pupils to be articulate by the end of Key Stage 2. **We recommend that the National Curriculum Review should consider how best to reflect its importance in the curriculum**.

Mathematics

The evidence we have received shows that the mathematics test is widely respected. 62% of respondents to the online call for evidence felt that the current mathematics Key Stage 2 tests were reasonably or very effective, as the following selection of comments indicates:

> "*These are the most reliable of the tests, by a distance. They provide a good range of question types and are a fair test of the mathematical knowledge and understanding that children at age eleven should reasonably be expected to have. They also test pupils' ability to apply what they know to some imaginative problems that children are unlikely to have encountered before. The threshold for each level is also about right: 79% + to achieve a level 5 reflects a sound grounding in the subject*."
>
> "*The maths tests generally give an accurate picture of pupils' attainment, are unambiguous and "child-friendly". However, more able pupils are disadvantaged as there is currently no level 6 test, and although teacher assessment at level 6 is possible, teachers are more likely to spend their time with pupils on the borderline of level 3 to 4 or level 4 to 5*."
>
> "*They require the application of linked mathematical concepts. Rote learning is not helpful – they do demand understanding*."
>
> "*The facts and figures nature of mathematics lends itself to tests*."

The Advisory Council on Mathematics Education (ACME) observed that the current mathematics tests are "*reasonably valid and reliable as measures of children's attainment in number, shape and space and data handling.*" However, ACME still questions whether a written test at the end of primary school is the best way of identifying the needs of children, and supports the development of robust teacher assessment to reflect the investigate and problem solving aspects of the subject[92].

While criticism of the current mathematics test was relatively limited, the amount of reading and the vocabulary expected of pupils is clearly a cause for concern:

> "*More time for children who struggle to answer all that they can... A child's maths understanding should not be spoilt by their reading speed*."
>
> "*Take away the emphasis on long worded questions that prevent children who are good at maths but not reading from accessing the tests*."
>
> "*Those children who struggle with reading cannot be effectively assessed in maths through the use of the test. Yes the test can be read to them but children do not raise their hand for every question.*"

A small number of respondents suggested that "*there is not enough emphasis on using the key skills through problem solving or using and applying skills.*" ACME also took this view, arguing that the fact that tests have insufficient focus on the use and application of mathematics means that pupils' classroom experience is consequently often narrowed. The current tests do include problem-solving questions, which may prompt some of the concern over the amount of reading in the test.

A few respondents questioned the need for a separate calculator paper: "*calculators are important tools for learning in the classroom, but the aim is to measure the children's mathematical ability and I think a single non-calculator paper would do that perfectly well.*" A small number questioned the mental arithmetic paper, including the time available for pupils to answer and the choice of the speaker's accent.

We have not received any evidence to suggest that there are significant issues with an externally-marked mathematics test. We recognise that it is relatively straightforward to create a valid and reliable test of mathematics, and we feel that the current mathematics tests achieve this. **We believe that it is legitimate to use a test to establish how well a pupil can perform a range of mathematical**

[92] ACME, 'Response to the *Progress Report* of the Bew Review into Key Stage 2 testing, assessment and accountability', submission to the Review (2011).

operations within a finite period of time. **We recommend that mathematics should continue to be subject to externally-marked testing**.

We acknowledge the concerns that results in the mathematics test should not be determined by ability in reading. **We recommend that in the development of future tests the amount of reading in the mathematics test should be kept under review, to ensure that weaker readers are not unfairly disadvantaged**. In addition, we believe that the current principle that questions should be placed in order of difficulty should be carefully adhered to in future mathematics tests.

We feel that it would be helpful for parents and secondary schools to receive detailed information on pupils' attainment within mathematics. Schools are currently required to make separate teacher assessment judgements for each attainment target in mathematics (use and application; number; shape, space and measure; handling data). **We recommend that summative teacher assessment in the mathematics attainment targets should be reported at pupil level to parents and secondary schools**.

Science

Although we have heard relatively little about the statutory assessment arrangements in science from the organisations and experts who presented evidence to us, the online call for evidence has revealed a diverse range of opinion. 51% of respondents felt that the current science assessment arrangements were reasonably or very effective. However, a significant minority (31%) felt they were not very effective or inadequate. The following quotes illustrate some of the main views:

> "*Teachers are free to teach Science as it was intended to be taught now, and do not have to teach to the test*."

> "*Now that pupils and teachers have been released from statutory tests, teachers are focusing more on the investigative aspects of science alongside knowledge and understanding and developing a range of assessment techniques including APP*."

While many respondents welcomed the move to use teacher assessment, many expressed concern at the lack of support for schools:

> "*The teacher assessment rather than test is a good step forward. The criteria for judging levels is not always detailed enough or with enough examples of what might be typical at each level. The level descriptions are designed to help with a summative judgement but need to be supported by more standardisation material and teachers need more support to moderate their judgements*."

> "*There are not enough guidelines for this subject since the removal of testing. It would be helpful to have a clearer idea of how we are to assess this huge, practical subject and what proof of assessment is necessary*."

Many respondents felt that the removal of statutory tests had significantly downgraded the place of science within the curriculum:

> "*Although imperfect and not properly addressing Attainment Target 1, the tests did test Science knowledge. The removal of the test did Science a disservice. It has dumbed down the subject.*"

> "*Now it is only a sample of schools and is not reported on a school by school basis the 'pressure' on schools to do well in science is much less. This is a shame. Removing Science whilst keeping Numeracy and Literacy has given a dangerous message to primary schools. In many, science is now one of the subjects you study after SATs preparation.*"

Science Community Representing Education (SCORE) (which comprises the Association for Science Education, Institute of Physics, Royal Society, Royal Society of Chemistry and Society of Biology) felt that the now-discontinued National Curriculum Tests in science "*reduced pupil motivation and enjoyment in science, greatly undervalued the professional judgement of science teachers, involved significant costs and acted as a barrier to innovation*"[93]. While they acknowledge the benefits of externally-marked testing, they feel they were outweighed by negative consequences for science teaching. They therefore support the present system whereby school report summative teacher assessment and a sample test is used to monitor national standards.

The Wellcome Trust cited their research[94] showing that, following the discontinuation of statutory tests in Wales "*children learned more about science and enjoyed the subject more without national tests.*" This study suggested that all groups of children preferred tests (though not National Curriculum Tests) rather than summative teacher assessment to determine their progress in science. The Wellcome Trust recommended that "*science assessment at KS2 should be embedded in normal science class work and should include the use of end-of topic, as opposed to end-of-year, testing. It should cover a range of sources of evidence from practical, oral and written work, and should focus on the understanding of science as opposed to knowledge recall.*"

We acknowledge that it is possible to create a valid and reliable test of scientific knowledge. However, we recognise that it is difficult to measure scientific enquiry (an important part of the curriculum) through an externally-marked test and that a focus on what can easily be tested risks distorting science teaching. As the current statutory assessment arrangements in science are relatively new, their effectiveness is not wholly clear, but we think the arguments for removing the test were justified. We therefore **recommend that pupil-level outcomes in science should continue to be based on summative teacher assessment.**

We believe it is important that national performance in science should continue to be monitored alongside schools' teacher assessment. **We recommend that sample testing in science should continue.**

We feel that it would be helpful for parents and secondary schools to receive detailed information on pupils' attainment within science. Schools are currently required to make separate teacher assessment judgements for each attainment target in science (scientific enquiry; life processes; materials and properties; physical processes). **We recommend that pupil-level summative teacher assessment in the science attainment targets should be reported to parents and secondary schools**.

In the long term, the Government should continue to seek feedback from schools and the science community as to the appropriateness and effectiveness of the current arrangements, particularly in view of changes to the curriculum. **We recommend that the current arrangements should be looked at again following the National Curriculum Review to ensure they are educationally appropriate for the new science National Curriculum**. We recognise that a specifically-designed sampling system could provide much more information than the National Curriculum Test papers currently in use. If the current arrangements are continued in the long term, **we recommend that a system of pupil-level sampling should be introduced**, because this would allow a greater coverage of the science curriculum than school-level sampling.

[93] SCORE, 'Key Stage 2 Testing and Accountability Review – Call for Evidence: SCORE response to Lord Bew's Review', submission to the Review (2011).
[94] Murphy, C., Kerr, K., Lundy, L., McEvoy, L., *Attitudes of Children and Parents to Key Stage 2 Science*, Wellcome Trust (2010).

Coherence between statutory assessment and the new National Curriculum

We are conscious that there are some inconsistencies between the way in which National Curriculum levels are currently assigned through National Curriculum Tests (based on specific mark schemes) and summative teacher assessment (which applies 'best fit' judgements). We acknowledge that this is, to an extent, an unavoidable consequence of maintaining two parallel approaches to assessment. However, these differences also reflect the way in which statutory assessment has been developed over time. In particular, the underlying Programmes of Study for each National Curriculum subject do not lend themselves to being used as an assessment framework. In English, for example, National Curriculum Tests are based on assessment foci derived from the National Curriculum. We feel that it would be helpful if there could be greater consistency between the National Curriculum, the expectations of summative teacher assessment and the way in which marks are assigned in statutory tests. **We would encourage the Government to seek greater coherence between the National Curriculum and its statutory assessment as an integral part of the design following the National Curriculum Review, without giving rise to a situation where statutory assessment can distort or narrow the curriculum**.

While statutory assessment will need to be designed to suit the new curriculum, we have had some discussions about how the two could support each other and drive pupils' learning. We would like to offer these suggestions which may be of value following the National Curriculum Review.

In the longer term, we feel it may be helpful for statutory assessment to divide into two parts. All pupils could be expected to master a 'core' of essential knowledge by the end of Key Stage 2, concentrating on the basic literacy and numeracy which all pupils require if they are to access the secondary curriculum. This 'core' could be assessed through a 'mastery' test which all pupils should be expected to pass (only excepting cases of profound Special Educational Needs), providing a high minimum standard of literacy and numeracy at the end of primary education.

We recognise the risk that this approach may lead to 'teaching to the test', may set an unhelpfully low ceiling on attainment and would not reflect pupils' progress. We would suggest two solutions. Firstly, it might be helpful to allow pupils to take 'core' tests in Years 4, 5 or 6 to ensure that able pupils are challenged. Secondly, we feel there could also be a separate assessment at the end of Key Stage 2 to allow pupils to demonstrate the extent of their knowledge and therefore to measure pupils' progress during the Key Stage. This assessment could be designed to identify the extent of pupils' attainment and understanding at the end of Year 6, spreading them out on a 'vertical scale' rather than being a pass/fail mastery test. Such an assessment should be as useful as possible to pupils, parents and teachers. It may be helpful for the results to report in greater detail than is currently provided by National Curriculum Test data, so they can identify more effectively the pupil's attainment in key broad aspects of a subject.

We feel the combination of these statutory assessments could ensure that all pupils reach a minimum standard of attainment while also allowing pupils to demonstrate the progress they have made – which would indicate the quality of the school's contribution to their education. It could provide a safety net in that all pupils should achieve a basic minimum, but would not impose a low ceiling on the able.

Chapter 4 – Delivery of testing and assessment arrangements

The majority of the evidence and feedback received by the Review, and consequently the majority of our recommendations, focused on accountability or statutory assessment and are therefore covered in the previous two chapters of this Final Report. As well as making recommendations for how the system should be improved, we have considered how to ensure the system of statutory assessment is as smooth and effective as possible. We have therefore made recommendations about how we envisage aspects of the system working in practice which we believe complement the recommendations in the first two chapters. We now turn to these.

This chapter is split into two sections; firstly, recommendations concerning delivery which we suggest should be implemented as soon as possible which cover moderation, secondary school transition, the timing of tests and on-screen marking. Secondly, we cover recommendations about changes that could potentially be introduced in the long term, which are dependent on the new National Curriculum and developments over time.

Cluster moderation to support professional development

As summative teacher assessment judgements of writing composition will be used to provide data for the accountability system, we have recommended that they should be moderated. This external moderation should demonstrate that the teacher assessment judgements are reliable and nationally consistent.

We recognise that this kind of external moderation will only go part of the way to developing teachers' assessment skills and allowing teachers to learn from each other. We feel that the moderation process should also include a focus on teachers building a shared understanding of educational standards.

We understand the value of groups of teachers from a range of schools (including secondary schools) meeting on a regular basis to build a shared understanding of educational standards and to discuss their assessment of pupils' work. **We would encourage schools to form clusters in this way to moderate teacher assessment judgements with the aim of learning from each other and developing the assessment skills of the teachers involved**. Many schools already participate in such networks, and we feel that other schools could benefit from adopting this approach.

The following quotes from the online call for evidence suggest this would be a worthwhile exercise:

> *"As a school that boycotted the SATs in 2010 my experience with cross-school moderation was very positive. Tests were taken in school, work was put alongside these results and a judgement was made through a professional discussion about the level of attainment."*

> *"Teacher assessment processes (not just the judgement) need to be explained and expanded. The moderation process is an excellent CPD opportunity when conducted in a rigorous and robust way. Time and opportunity needs to be given to schools to work collaboratively on making the teacher assessment process effective."*

In line with the philosophy of the Schools White Paper of greater freedom and autonomy for schools, we do not wish to prescribe how this moderation should happen. We would suggest simply that schools cluster together as they see fit,

perhaps around an outstanding school, and decide on the moderation activity which best enables teachers to learn from each other and develop their assessment skills.

Transition to secondary school

Feedback to the Review suggests that providing information about pupils' attainment on transition to secondary school is critically important, but that secondary schools currently make limited use of the statutory test and teacher assessment data which they receive.

We realise that the current challenges around transition to secondary school cannot be solved by recommendations about statutory assessment. However, we feel that we can help reduce some of the sources of friction which lead to a limited use of the information primary schools provide.

41% of online call for evidence respondents said that one of the most important purposes of the tests should be to provide information about pupils' attainment at the time of transfer between schools.

However, secondary school respondents have expressed concern that end of Key Stage 2 results are not always a suitable proxy for the attainment of pupils on entry to Year 7. Many secondary schools and head teachers have given evidence about the perceived disparities between the level gained at the end of Year 6 and pupils' subsequent performance at the start of Year 7. Many secondary schools are concerned that over-preparation in Year 6 means that test results may not give an accurate reflection of a pupil's actual attainment.

The following quote gives an example of the concern about transition to secondary school:

> *"I find it alarming that most secondary schools re-test children within their first six weeks of the autumn term. This is over testing young children and a firm decision must be made as to when assessment should be made - either at the end of KS2 or at the beginning of KS3 – not both. We need to develop trust between staff working within the transition to support the transfer of children within establishments."*

Research evidence shows that secondary schools make widespread use of Cognitive Abilities Tests (which are designed to predict ability, regardless of previous attainment) and other internal assessments[95]. The wide range of commercially available tests that are bought in by secondary schools shows the level of demand for additional information. This may suggest that they doubt the usefulness or accuracy of National Curriculum Test results, or that they see the value of triangulating with a different form of assessment. We recognise that secondary schools will wish to reach their own teacher assessment judgements on pupils once they have spent some time in Year 7 and will draw on a wide range of information to do so. We wish to ensure that statutory assessment data from the end of Year 6 (both tests and summative teacher assessment) is as timely and useful as possible.

We believe that two recommendations which we have already made will make statutory assessment more useful to secondary schools. Firstly, we have recommended that what is reported from statutory assessment should be more detailed than at present. In particular, we have recommended that English should be broken down into separately-reported individual components (reading, writing and speaking and listening) and that teacher assessment levels in each mathematics Attainment Target should be reported, alongside an overall teacher assessment level and test results. We feel this will have particular benefits for parents, but also for

[95] Kirkup, C., Sizmur, J., Sturman, L. and Lewis, K., *Schools' use of data in teaching and learning,* NFER (2005).

secondary schools as Year 7 teachers will be able to get a better understanding of their new pupils from the start.

Secondly, we have recommended that the date for submitting teacher assessment results should be brought forward to increase the emphasis on teacher assessment. We believe an additional advantage of this recommendation is that secondary schools will receive the teacher assessment data earlier in the summer term before the new intake arrives, which will allow Year 7 teachers longer to use the information for planning purposes.

Given the improvements to the information secondary schools will receive, we encourage secondary schools to make wider use of the pupil-level data available from Key Stage 2 to support transition of new Year 7 pupils. However, we recognise that even if statutory assessment data is provided in greater detail, earlier in the year, and used effectively by all secondary schools, it is still only part of the information secondary schools need. Primary schools will be able to send additional information based on their knowledge of pupils.

Given the greater focus on teacher assessment information, we feel there is potential in encouraging cross-phase moderation of Year 6 pupils' work. We believe Year 7 teachers should be involved in the moderation of teacher assessment judgements of Year 6 pupils' writing composition work in particular as moderators themselves. We encourage secondary schools to engage with this approach and also recommend that the Government should consider what incentives can be put in place to encourage Year 7 teachers to join in moderation exercises with Year 6 teachers designed to support professional development.

We see cross-phase moderation as a way of supporting continuing professional development, building a shared understanding of the importance of assessment, and delivering more meaningful data for secondary schools. We believe this approach would allow Year 7 teachers to develop a better understanding of the standards of attainment at Key Stage 2, and help develop their trust in the information that primary schools provide.

Transition from Key Stage 1 to Key Stage 2

Feedback to the Review has suggested that similar transition problems can occur to an extent between infant and junior schools. We believe these problems could be partly tackled through cross-phase moderation where it does not already happen. **We believe the same principle of encouraging cross-phase moderation should apply to infant and junior schools, and we encourage moderation of Key Stage 1 teacher assessment judgements involving both Year 2 and Year 3 teachers from infant and junior schools.** Cross-Key Stage moderation of Key Stage 1 teacher assessment judgements gives Year 3 teachers a better understanding of their new intake. We welcome the fact that this is already common practice in many schools.

In the Accountability chapter we recommended the moderation process at Key Stage 1 is developed further to be more consistently rigorous. We believe encouraging cross-phase moderation between infant and junior schools complements this recommendation. This will help ensure that Key Stage 1 pupil-level data is robust and that Year 3 teachers feel confident in making wide use of it to understand their new intake.

Timing of tests

The timing of tests has been debated. Opinion is mixed, particularly amongst those who responded to the online call for evidence. 31% advocated shortened tests at more than one point. 27% suggested retaining tests in Year 6. 18% recommended tests at the start of secondary education. 30% selected 'other'.

The following quotes from the online call for evidence show the range of opinion in this area:

> *"Secondary schools already carry out tests at the start of Y7; however, whilst KS2 schools are still judged on these tests, I do not think that moving them to Y7 is fair. We already see pupils who do well within a primary school environment struggle with the transition to secondary school. For them to complete tests to show their progress during KS2 at the start of what can be a difficult year for transition would not be fair to the pupils or the KS2 feeder school."*

> *"Current time is great. Definitely not at the beginning of Key Stage 3 as children at this time are coping with transition and testing now would be unfair, stressful and would not allow them to have a successful start to secondary education."*

> *"Earlier in the Key Stage, or at least earlier in Year 6 would allow us to reflect and act on results to ensure future planning addresses pupils' needs. By the time we get the results our children have left for secondary school."*

> *"If they have to be taken then they should happen at the end of the final term not the beginning as they do now, children can change and learn more in the remaining weeks. After the summer holidays many children will have slipped back due to the lack of practice and have enough to think about with starting a new school."*

It is worth noting at this stage that relatively few respondents have given evidence to the Review concerning the timing of tests. Advocates of moving tests to the start of Year 7 argue that it would ease pressure in Year 6 and change teachers' practice from short-term 'drilling' to longer-term preparation for the new school year. This could encourage a move away from the mindset of short-term learning and would make it harder for pupils to be awarded levels where they are not secure.

However, many stakeholders have questioned the impact of conducting external tests early in Year 7. A sample of heads interviewed by QCDA overwhelmingly felt that tests should be administered within the phase being held to account. ASCL, representing secondary heads, felt it was not appropriate for secondary schools to be responsible for administering tests which would be used to hold primary schools accountable. We have also heard concerns that the well-attested 'Year 7 dip' may make it difficult to interpret results taken after the summer holiday.

Some respondents have suggested that tests should happen later in the summer term to maximise teaching time. However, when QCDA consulted in 2009 on moving the 2011 test dates to June, many schools and local authorities raised concerns, including the impact on educational trips and out of classroom learning opportunities, regular summer sports events and secondary transition arrangements.

Professors Baird and Elwood suggest that a national test taken towards the end of Year 5 or at the beginning of Year 6 could act as a measure of how well pupils are doing and what they need to do to improve.

As these views indicate, opinion on the timing of tests is mixed, but we have focused most on considering whether a move to Year 7 would be the right solution as this is the most substantial suggestion for change and would transform the current system if implemented.

Based on feedback from primary and secondary head teachers, local authorities and markers (collated by QCDA), we believe that it is an operationally feasible option. However, there is no evidence that it would reduce test preparation. Indeed, many primary head teachers felt that the period for test preparation would increase and

potentially extend over the summer holiday. Secondary heads felt that moving the tests could even increase the practice of testing on entry, as access to Key Stage 2 test data would be delayed.

Concerns were also raised about the possible effects on teaching in Key Stage 3, and a potential delay in teaching the Key Stage 3 programmes of study, as schools might maintain focus on Key Stage 2 teaching until the tests. Conversely, there were also concerns that teaching the Key Stage 3 programmes of study prior to tests in Year 7 would lead to pupils being less familiar with the Key Stage 2 programmes of study (on which the tests would be based), particularly after the summer break.

In addition, several concerns were raised over the possible effects on the performance of pupils in the tests, in particular that the long summer break from teaching and the impact of a new school environment could lead to lower results. Head teachers felt that the detrimental effects of moving tests to Year 7 would be most marked for vulnerable pupils who are more likely to experience difficulty in transition, and by pupils with English as an additional language, who would be less likely to communicate in English during the summer break.

The evidence and feedback to the Review suggests that changing the timing of end of Key Stage 2 tests to the beginning of Year 7 is a feasible option, but we believe is not the best solution to the problems with the current system. We therefore recommend that the timing of tests remains as it is.

We have heard a range of feedback and opinion about when in the summer term the Year 6 tests should take place. Some argue that tests are too early, and do not allow for the full year to be used for teaching and learning. Others say that they value the time after test week when Year 6 are relieved of the pressure of testing and can focus more on other areas of the curriculum, and that they would oppose any move to place the tests later in the school year.

Based on the evidence and feedback we have received, we do not believe that there is a compelling argument to move statutory tests to significantly earlier or later in the summer term.

On-screen marking

The concept of on-screen marking, whereby scripts are collected, scanned and marked online, has become a routine part of many test processes. It is currently used for marking the science sample test, but is not a feature of National Curriculum Tests. QCA's evidence to the 2008 Sutherland Inquiry observed that on-screen marking *"is not only more accurate and reliable than manual marking, but faster. While not all types of examination are suited to onscreen marking, it is now a proven and common approach*"[96]. In its review of the 2010 National Curriculum tests, Ofqual noted "*the successful introduction of onscreen marking for key stage 2 science sampling made the overall marking process more efficient*"[97].

We believe that on-screen marking would bring about improvements to the current system. These include greater speed of marking, allowing results to be returned to schools more quickly; facilitating the return of item-level data, showing how well pupils did on particular questions and areas of the curriculum; greater security in the marking process, reducing the need for scripts to be physically sent to and from markers; greater specialisation of markers, who can focus exclusively on sections of a test paper rather than the whole paper; and greater potential for double marking, which would improve reliability.

[96] QCA submission to the Sutherland Inquiry, 12 September 2008, in Sutherland, S., *The Sutherland Inquiry*, (2008).
[97] Ofqual, *National Curriculum Assessments Review Report: 2010 Key Stage 2 tests*, (2011).

However, we recognise the need for confidence in any new ICT system, particularly given the number of pupils and schools that would be involved.

We believe that on-screen marking should be considered for other Key Stage 2 tests. We recommend that the Government should learn from the evidence from science sample tests and plan what further trialling is needed with the aim of moving to a full rollout of on-screen marking.

Potential long-term changes

We have heard a great deal about various alternative approaches to statutory assessment and their advantages and drawbacks, which we cover in this section of the Final Report. We believe caution is needed before introducing any of these new approaches, so we do not wish to make any detailed recommendations, but we believe their potential means that they should be considered in greater detail in the future, particularly alongside the new National Curriculum. We believe the approaches we go on to cover could improve the delivery of the statutory assessment system we set out in the previous chapter.

Computer-administered testing

A number of respondents have suggested that National Curriculum Tests could be administered using computers rather than the current 'pencil and paper' format. The following quotation from the online call for evidence summarises the potential benefits:

> "*Online testing would be much more cost-efficient and less stressful for children as most are very computer literate. It would be environmentally friendly and would be cheaper and much easier to mark.*"

We recognise that computer-administered testing could greatly simplify test delivery. In particular, it may be possible for tests to be marked automatically, although some types of question may still require a human marker. This could reduce the costs of marking considerably and allow schools to receive results shortly after the test was completed, broken down question by question.

There are examples of online testing in other countries from which we can learn. Tim Oates highlighted a test developed by Massachusetts Institute of Technology (MIT) in physics as a good example of the potential of online testing. Sir Michael Barber similarly highlighted experiments in computer-driven assessment in the US and Hong Kong. In New York the 'Wireless Generation' company provides online testing materials, which they mark and return to schools together with comparisons against schools with similar results.

However, we accept that computer-administered testing needs to be approached with caution. Tests may need to be designed differently, making use of different kinds of question (for example through greater use of multiple choice answers). Most importantly, it would be essential that the technical delivery of computer-administered testing was wholly robust, particularly if tests were to be administered to all pupils at the end of Year 6 simultaneously. The Government would also need to be satisfied that all schools had access to sufficient computers for their whole Year 6 cohort to take the same computer-administered tests at once.

Considerable as these challenges may be, we feel that the Government should consider the potential of computer-administered testing in the long term. There will need to be thorough piloting and preparatory work over a number of years. **We recommend exploration and piloting of computer-administered testing**.

Computer adaptive testing

Some respondents have pointed to the potential of one specific type of computer-administered test – computer adaptive testing. Pupils are presented with their own personalised test according to their level of performance. Questions are selected from a bank of questions using item response theory algorithms, with the aim of establishing an exact level of performance. For example, if a pupil gets questions of a particular difficulty consistently right, harder questions will be presented until it is clear the pupil's ability threshold has been reached. This approach typically uses a multiple choice format, although it can be more complex. Results are generated instantaneously. However, it is worth noting that, because of the adaptive way in which questions are generated, pupils will not sit identical tests. While individual pupils' test results may be accurate, there may be challenges in comparing results. There are several arguments in favour of adaptive testing, which is already in use in primary schools in other countries, and many primary schools in Northern Ireland use adaptive tests. The proponents of adaptive testing point to significant efficiencies, arguing that a test need only be half as long as a typical National Curriculum Test in order to provide the same degree of precision. Adaptive testing can give both a score for summative assessment and more detailed information for formative assessment.

However, we are aware of the potential limitations of the system. There may be practical (and security) issues in ensuring that all children sit the test at a particular time. This may be exacerbated by whether or not the school has sufficient IT facilities to accommodate all pupils in the year group. The use of computer adaptive testing also assumes a certain degree of IT literacy on the part of the child, which may disadvantage certain groups of pupils.

We believe the potential of computer adaptive testing should be explored further, including the relative suitability of the system for assessing specific subjects, with a view to exploring the possibility of introducing in the long term.

Testing when ready

A key criticism of the current Key Stage 2 tests is that pupils' knowledge and skills over a four-year Key Stage is assessed via tests in a single specified week in May. Some critics have raised concerns that this approach causes stress for pupils, particularly those working at the lower end of a spectrum, and may have unfair implications for schools, whose overall results may be affected if for example a highly-performing pupil is absent on test day. In addition, criticism suggests there is little incentive to challenge the more able children, who may well be working at level 5 at an earlier point in the Key Stage or year.

We believe that our earlier recommendations address these issues. However, we also recognise the benefits of a system based on the principle of 'testing when ready'. The proponents of such an approach argue that it would allow each pupil to be entered for statutory tests when he/she is ready, and then able to move on to more advanced learning. We believe that it would be possible for a statutory 'testing when ready' system to meet the statutory assessment purposes we have specified.

However, we are not convinced that moving to a 'testing when ready' approach is the best way of achieving the purposes of statutory assessment under the current National Curriculum. **We suggest that the principle of 'testing when ready' should be considered in the future following the National Curriculum Review.** We believe that the principle of 'testing when ready' may fit well if computer-administered testing is introduced, making it easier for each pupil to sit his/her own personalised test at any point in time when teachers deem him/her to be ready.

We have heard evidence and feedback about Single Level Tests, which allowed pupils to be tested on whether they had achieved a particular National Curriculum

level at any point in the Key Stage. Single Level Tests did not offer a pure 'testing when ready' approach since they could only be taken at two fixed points each year, rather than at any point in a year. They relied on the judgement of teachers about whether pupils between Year 3 and Year 6 were ready to be entered for a particular level. The 60-minute Single Level Tests each covered a single National Curriculum level (levels 3–6), providing a link between formative teacher assessment and summative testing.

As part of the Making Good Progress pilot, which ran in 10 local authorities from 2007-2009, pupils in pilot schools could be entered for Single Level Tests in June and/or December as well as (but not instead of) the National Curriculum Tests. In the third year only, the mathematics Single Level Test was used in an accountability context (i.e. as a replacement for the mathematics National Curriculum Test). The pilot ended in September 2010.

The pilot schools reported that this approach to testing contributed to a broader and more balanced curriculum across Key Stage 2, particularly in Year 6. However, some heads from the pilot schools did feed back in oral evidence to the Review that excessive test preparation might still be an issue with Single Level Tests. The pilot schools reported that combined use of Single Level Tests and Assessing Pupils' Progress (APP) materials impacted positively on teaching. Tracking of pupils' progress improved as a result, benefitting most Key Stage 2 pupils, particularly the more able.

Head teacher representatives from the pilot who gave evidence to the Review commented positively on the use of Single Level Tests in their schools. They felt that the tests quality-assured teacher assessment and therefore increased teachers' confidence. We hope that our recommendation that teacher assessment judgements should be submitted ahead of test results (as explained in the Accountability Chapter) will help to build teachers' confidence in a similar way. Pupils felt less intimidated by taking a test aimed at their level rather than a test covering several levels, especially as they were aware that in most cases there would be a second opportunity to take the test.

However, some feedback to the Review has suggested that the Single Level Tests concept proved complex, and we have heard criticism from a range of stakeholders (though not those involved in the pilot) that, in their view, a completely effective model had not been reached even after three years of piloting. There has been support for a system of 'testing when ready', but we believe such a system would benefit from allowing more frequent testing opportunities than the two test windows each year provided by Single Level Tests. This would require considerable further development

Based on the mixed feedback that we have received about Single Level Tests, we are not convinced by the model that was piloted and we therefore do not recommend that the model is continued. However, we believe that, if we move to a 'testing when ready' approach in the future, we can learn lessons from the pilot and in particular the benefits it has achieved.

Annex – Stakeholders who have submitted evidence to the Review

Oral evidence

We are grateful to the following individuals and organisations who presented oral evidence to the panel, together with a written summary.

Association for Achievement and Improvement through Assessment (AAIA)
Advisory Council for Mathematics Education (ACME)
Professor Robin Alexander
ARK Schools
Association of School and College Leaders (ASCL)
Association of Teachers and Lecturers (ATL)
Professor Jo-Anne Baird
Professor Sir Michael Barber
Julian Barrell, Director, Simply Efficient Ltd
Professor Margaret Brown
Cambridge Primary Review
Professor Richard Daugherty
Professor Janette Elwood
GL Assessment
Professor Wynne Harlen
Bill Holledge, Culloden Primary School, Tower Hamlets
Dr. Tina Isaacs
Warwick Mansell
Dr Christine Merrell
National Governors' Association (NGA)
National Association of Head Teachers (NAHT)
National Association of Schoolmasters/Union of Women Teachers (NASUWT)
National Foundation for Educational Research (NFER)
National Union of Teachers (NUT)
Dr. Paul Newton
Tim Oates
Ofsted
Pearson UK
Qualifications and Curriculum Development Agency (QCDA)
Steiner Waldorf Schools Fellowship
Lord Sutherland
Professor Peter Tymms
Professor Dylan Wiliam

Heads and teachers

We are particularly grateful to the heads of the following schools who took the time to meet the Panel to offer their feedback and share their experiences.

Jeremy Bird, Head Teacher, Greswold Primary School, Solihull
Lynne Bruce, Brookside School, Leicestershire
Barbara Coates, Head Teacher, Little Hallingbury CE Voluntary Aided School, Essex
Val Cobb, Hawkes Farm Primary School, East Sussex
Charles Daniels, Sacred Heart Catholic Primary School, Liverpool
Tony Draper, Head Teacher, Water Hall Primary School, Milton Keynes
Sian Fenton, Head Teacher, Shelf Junior and Infant School, Calderdale
Karine George, Head Teacher, Westfields Junior School, Hampshire
Katherine Leahy, Cam Hopton CE Primary School, Gloucestershire
David Linsell, Ratton School, East Sussex
Tony Markham, Head Teacher, The Herne Junior School, Hampshire
Tony Newman, Head Teacher, Stanley School, Wirral
Debra Okitikpi, Head Teacher, Edward Wilson Primary School, Westminster
Kevin Parfoot, Head Teacher, Purbrook Junior School, Hampshire
Paul Williams, Head Teacher, Shaftesbury High School, Harrow
Peter Wilson, Old Bexley CE Primary School, Bexley
Kate Utting, Head Teacher, Horndean CE Junior School, Hampshire

Written evidence

We would like to thank all those who have provided written submissions to the Review to put forward detailed evidence and feedback.

Association of Directors of Children's Services (ADCS)
Assessment and Qualifications Alliance (AQA)
Centre for Policy Studies
Chartered Institute of Educational Assessors (CIEA)
Department for Children, Education, Lifelong Learning and Skills, Welsh Assembly Government
Girls' Day School Trust (GDST)
General Teaching Council for England (GTCE)
Colin Green, Director of Children, Learning and Young People, Coventry City Council
C J R Luckin, Head Teacher, St Andrew's CE Primary School, Steyning
National Education Trust (NET)
National Primary Headteachers' Association (NPH)
Ofqual
Alison Peacock, Head Teacher, The Wroxham Primary School, Hertfordshire
Alastair Pollitt
Renaissance Learning
Professor Colin Richards
Professor Pam Sammons
Alan Simpson, Head Teacher, Seaton Primary School, Devon
Science Community Representing Education (SCORE)
The Tiptree and Stanway Primary Schools Consortium, Essex
Roger Titcombe
Professor Harry Torrance
Wellcome Trust

Online call for evidence

An online call for evidence launched on 25 November 2010 and closed on 17 February 2011, attracting **3,940** responses. Of these, **2,386 (61%)** were primary head teachers; **900 (23%)** were primary school teachers; **156 (4%)** were parents/carers; **124 (3%)** were governors; and **59 (1%)** were pupils. The remaining **8%** of respondents comprised: education professionals; local authorities; secondary school teachers and head teachers; teaching unions/professional associations; and others. We have also published a *Call for Evidence Report*, which summarises the main findings from the Review's call for evidence.

84% of responses to the online call for evidence were from primary school head teachers and teachers (over 3,000 in total). This is a very large proportion of the responses, particularly in comparison with other important groups of stakeholders such as governors, parents/carers and secondary school head teachers and teachers.

A significant number of educational professionals and associations chose to submit detailed evidence and feedback to the Review through the online call for evidence.

This publication is available for download at www.education.gov.uk