

Rt Hon Michael Gove MP
Secretary of State for Education
Department for Education
Sanctuary Buildings
20 Great Smith Street
London SW1P 3BT

1 October 2013

Dear Michael

Response to consultation on primary assessment and accountability

Enclosed is the response of the British Educational Research Association to the consultation launched in July on primary assessment and accountability under the new national curriculum.

Given the vital importance of developing a defensible and fair national system, we decided to take the unusual step of reconvening the influential UK Assessment Reform Group, which remains highly regarded nationally and internationally. The ARG formally retired in 2010 after more than 20 years of voluntary work to bring assessment research evidence to the attention of policy makers. Its previous outputs can be found in its archived website:

<http://webarchive.nationalarchives.gov.uk/20110809101133/assessment-reform-group.org/>

The enclosed response from the group is therefore a well-considered, collective effort from the ARG, endorsed by BERA. We hope that you will find it of interest to your deliberations. When you, and David Laws, have had time to read it, the group would be happy to meet you to discuss it further.

Yours ever,

Mary

Professor Mary James
Vice President BERA
Convenor ARG

Email: mej1002@cam.ac.uk
Tel: 07714 652 809



**A response to the DFE's 2013 consultation on:
*Primary assessment and accountability
under the new national curriculum***

1. Introduction

- 1.1. We share the Government's aim (para. 1.1 of the consultation document) for the school system in England to match up to the highest standards internationally with clear, ambitious expectations for the attainments of **all** pupils.
- 1.2. We also agree that when designing the new system of assessment in primary schools we can learn from the experience of other high performing jurisdictions. The key features of such systems have been summarised in a recent OECD report: *Synergies for Better Learning: an international perspective on evaluation and assessment* (OECD 2013). This sets out policy directions that are highly relevant to the consideration of the DfE's proposals for primary education in England. The key features of the report offer criteria against which the current proposals for England can be judged. Our response, therefore, is made with these in mind.
- 1.3. Derived from reviews of assessment and evaluation practice in 28 OECD countries, including the most highly performing jurisdictions from which we are keen to learn, the recommendations of most relevance are to:
 - i. integrate student assessment and school evaluation in a single framework which 'articulates ways to achieve the coherence between its different components';
 - ii. align assessment and evaluation with educational goals and learning objectives set out in the curriculum;
 - iii. design the accountability uses of evaluation and assessment in ways that minimise undesirable effects;

- iv. use measures of performance that are broad enough to capture the whole range of student learning objectives;
 - v. focus on improving classroom practices and build on teacher professionalism. (To 'optimise the potential of evaluation and assessment to improve what is at the heart of education – student learning – policy makers should promote the regular use of evaluation and assessment results for improvements in the classroom.');
 - vi. place the student at the centre, fostering engagement in learning through using formative assessment strategies.
- 1.4. We endorse the overall **purpose** of (summative) assessment in primary schools (consultation document, para. 4.1), defined in legislation as 'to ascertain what pupils have achieved in relation to the attainment targets' and we acknowledge the argument in Lord Bew's report that there are three main **uses** of data derived from such assessments:
- i. *holding schools accountable* for the standard of attainment and progress made by their pupils and groups of pupils;
 - ii. informing parents and secondary schools about *the performance of individual pupils*; and
 - iii. *enabling benchmarking* between schools; as well as *monitoring performance* locally and nationally.
- 1.5. However, we wish to reiterate the important point made in our 2009 commentary, *Assessment in Schools: Fit for Purpose?*, that multiple **uses** can subvert the overall **purpose** if they have negative unintended consequences. We wrote then (Mansell, James and the ARG, 2009, p.7):
- Assessment information has become a proxy measure that is supposed to facilitate judgments on the quality of most elements of our education system: its **teachers, head teachers, schools, support services, local authorities** and even the **government** itself. This represents a fundamental change from the situation even 20 years ago, when test and examination results were predominantly meant to serve as indicators of what a **pupil** knew and understood of a subject.*
- 1.6. That said, our response to the current proposals is organised around the three main **uses** identified by Lord Bew, although we adopt the sequence: individual reporting; school accountability; benchmarking and monitoring. Each of the uses will need clear definition and a national framework even if, as with individual reporting of attainment and with some aspects of accountability, they are most appropriately determined at the local level.
- 1.7. We agree that there should be a distinction between ongoing formative assessment, to be developed and used within schools, and statutory summative assessment (consultation document, para. 1.6).

- 1.8. However, we suggest that the statutory elements in the national system should be seen as comprising *two parallel and complementary sub-systems* of summative assessment: national tests and teacher-based assessments.
- 1.9. **Statutory test-based assessment** should be planned and managed nationally with the data from the tests also used locally i.e. by schools and local authorities. This sub-system is justified mainly in terms of national monitoring and benchmarking but it would also contribute to how schools are held accountable.
- 1.10. **Statutory teacher-based assessment** should be planned and managed by schools within a national framework. Data from these sources would be used primarily to provide essential information on the progress of individual pupils.
- 1.11. Assessment for each of the three uses must be curriculum-related. We argue that this requires a national framework to guide summative assessment practices.
- 1.12. Assessment expertise also needs to be developed and effectively supported within schools if the new national system is to achieve its goals.

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

- 2.1. In principle, individual reporting mechanisms should be designed to reflect the kind of information needed by those who will use the results: specifically **parents** and **'receiving' teachers** (responsible for the next stage of a child's educational experience).
- 2.2. In the following paragraphs we treat separately the test-based and teacher-based sub-systems of summative assessment mentioned above.
- 2.3. **Teacher-based assessment** (both statutory and non-statutory) can serve formative and summative purposes. The consultation document makes a 'clear separation' between formative assessment and summative assessment (para 1.6). We agree that the approach to formative assessment should be the responsibility of the school; the approach to statutory summative assessment should be prescribed by Government. Both need to be referenced to the curriculum.
- 2.4. In relation to formative assessment we underline its strong role in raising the standard of attainment of every pupil. It is well established by research that this is of particular benefit to the lower

attaining pupils and therefore plays an important role in ensuring that all pupils achieve at the highest level possible, and that schools' attainment is above floor standards.

- 2.5. In relation to statutory teacher-based summative assessment for reporting purposes we urge that this should be a required part of a coherent national assessment framework, as advocated by the OECD. Summative teacher assessment can be based on many observations of an individual pupil's performance, across the curriculum and over time. In addition, it is possible to strengthen the comparability and credibility of a teacher's own assessments by internal (to the school) and external moderation. Although these processes take time to develop and embed, and also take time to implement on an ongoing basis, moderation activities constitute some of the very best professional development. They provide unique opportunities for teachers to discuss, and continually refine, their understanding of standards of quality and progression within learning domains. These understandings are fundamental aspects of pedagogical content knowledge, crucial for effective teaching and for understanding what constitutes high expectations - and appropriate cognitive demand - in specific subject areas.
- 2.6. Statutory teacher-based assessment should be used for reporting individual pupils' achievements to parents, pupils' next teachers in the primary school and secondary school at transfer, although the nature of these reports may differ for parents and receiving teachers. The new national curriculum programmes of study set out what teachers should teach pupils to know, understand and do. These represent criteria and standards that teachers can use to assess and report pupils' achievements. They also, importantly, embody expectations of progression in learning in the way that they are set out, by key stage or year-by-year.
- 2.7. **Qualitative reports (i.e. using natural language rather than levels, grades or scores) on individual pupils should include attainment in all subjects and should be achievement-related. They should indicate what the pupil knows and can do in relation to the national curriculum requirements identified in the programmes of study and attainment targets.** Achievements in other (non-statutory) subjects chosen by the school should also be included in the reports.
- 2.8. Qualitative reports should not be translated into levels, grades or scores for central collection or for managerial/accountability purposes, because this will distort the process and subvert the primary purpose of individual reporting.
- 2.9. We propose therefore that statutory teacher-based assessment should be the main source of information on individual pupils, and that there should be **no requirement to report the results of**

teacher assessments in summary form to a national database for use in national monitoring. National test results should be sufficient for benchmarking and accountability purposes.

- 2.10. Parents will, of course, wish to be informed about how individual pupils performed on **test-based assessments**, as will teachers. This is best done through the use of scaled scores. As noted in the next section, there is nothing to be gained – and indeed there is much to be lost – by the use of deciles. These create artificial boundaries and risk being treated in the same way as the existing ‘levels’ to label pupils, with a consequent de-motivating impact on both high as well as low scorers.
- 2.11. The proposal that a cut score of 100 on the national tests should signify that a pupil is ‘secondary ready’ is likely to be equally damaging. This arbitrary hurdle, to separate the ready from the unready, bears no relationship to the concept of ‘ready to progress’ as defined in the report of the National Curriculum Review Expert Panel (para. 8.7). In this report ‘ready to progress’ referred to ‘all pupils having an adequate understanding of key elements [of the curriculum] prior to moving to the next body of content’. This has to be judged qualitatively and cannot be ascribed to a single cut score.
- 2.12. **It is our view, therefore, that the proposed use of scaled scores will provide sufficient information to enable parents and teachers to compare the attainment of individual pupils on national tests with the cohort average.** The use of deciles and a ‘secondary ready’ cut score, where a single mark can make all the difference, is unnecessary and will have adverse effects.
- 2.13. We need to reach a point very soon where primary pupils, their parents and teachers, are less focused on abstract levels, grades or scores, and more on whether pupils have learned the substantive facts, concepts, skills and attitudes that will equip them for future learning.

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

- 3.1. It is important to note that the Written Ministerial Statement of 17th July 2013 stated that ‘statutory assessment in core subjects at the end of key stages is designed primarily to enable external accountability’. We accept that national curriculum tests, as proposed, provide data in a form suited to this specific use. However, it has to be recognised that the tests in English and mathematics can only sample the programme of study of a key stage in very limited ways. Thus, validity in relation to the whole subject will be limited. While the reliability of results for monitoring the performance of a cohort may be robust - errors can cancel out

one another - the reliability of results for a single pupil will inevitably be fragile, even at KS2.

- 3.2. At KS1, we recommend not relying on results from a single test as a baseline for judging the subsequent progress of individual pupils. Even tests that appear to be reasonably reliable, by psychometric standards, will be inaccurate for a substantial percentage of pupils. The attainment (and subsequent progress) of *individual pupils* should be judged holistically in relation to a *range* of assessments. When results are aggregated across pupils in a school, they become more reliable. However, this cannot eliminate certain kinds of bias. Moreover, when results are used for accountability purposes, and teachers have some say over those results, there is an extreme and perverse incentive for teachers to bias their judgements (the temptation is to deflate baseline judgements, in order to inflate progress scores). Experience has repeatedly shown that a substantial proportion of teachers, whether implicitly or explicitly, succumb to this temptation to bias results. This is a major risk to the accuracy of results, and to the defensibility of using such progress measures for accountability purposes. We support the continued current arrangements for teacher assessed tests and tasks at KS1 for reporting to parents.
- 3.3. The problems of dependability (relating to the interconnected properties of reliability and, especially, validity) of test results would apply to any attempt to create a new baseline test at the beginning of reception, so we vigorously oppose this idea. Once again, the principal objection is the incentive for teachers to bias (deflate) the assessment outcome, whilst administering the baseline assessment.
- 3.4. The revised EYFS Profile is a useful assessment at the early years stage. But the breadth of its compass, which is an advantage for the purpose of identifying pupils's needs, is a disadvantage for the purpose of providing a baseline for narrower national tests in reading and mathematics at KS2. Whilst certain of its sub-scales could be used, they are unable to discriminate sufficiently to provide a good baseline measure alone. Moreover, as an accountability measure, it would be undermined by pressure on teachers to bias their judgements. We certainly would not want to see the EYFS Profile made non-statutory in favour of a new, narrow 'baseline check'.
- 3.5. We acknowledge the widely-held view that schools cannot be judged or compared purely in terms of outcome measures at the end of KS2 because pupils in different schools will have started from different baselines of attainment in reading and mathematics. However, to assess young children effectively requires support from teachers, even if this is simply to support task administration. This again exposes the assessment to the incentive to bias results. For this reason the results of early years assessments at age 4, 5 or 7 ought not to be included in value-added calculations for judging and

comparing schools. The progress that a pupil makes in passing through the school will be evident in the year-by-year qualitative reports.

- 3.6. A more defensible alternative for the purpose of benchmarking is to compare the national test results of schools with similar background variable profiles, such as free school meal entitlement (or another proxy for SES), SEND, English as an additional language, and children from different ethnic backgrounds.
- 3.7. On the matter of school accountability, the Bew report reached the important conclusion that ‘a broader range of published data [than test results] would help ensure schools are held accountable in a fair way’ (Bew Report, page 24). With this in mind it is important to go beyond accountability based on test data, which is the focus of most of the proposals in the consultation document.
- 3.8. To measure attainment at the school level, the use of aggregated scaled scores from KS2 tests avoids the negative effects of school targets that focus on particular levels. This is therefore to be welcomed. New targets need to be more inclusive of all pupils within schools. Ultimately, though, what schools are rightly accountable for is not an aggregated overall level of attainment, but how effective they are in developing the learning of pupils with different needs and backgrounds. This will require more sophisticated modelling (e.g. multi-level modelling) of how different factors affect attainment at KS2, in order to show how a school has created positive trajectories for pupils in different categories. Ofsted should be responsible for preparing and reporting these analyses, and for synthesising them with additional information on school effectiveness from other available data and from school inspections.
- 3.9. As noted earlier, we consider that it is not helpful to publish any end of key stage teacher-based assessments. The published results should be based on national curriculum tests only, and as three-year rolling averages, as recommended by the Bew report.
- 3.10. Finally, in relation to school-level accountability, we wish to draw attention to the fact that high-performing jurisdictions, such as Finland, Singapore and Hong Kong, do **not** publish performance tables of test or assessment results aggregated at school level.

4. *Enabling benchmarking between schools; as well as monitoring performance locally and nationally*

- 4.1. We recognise the utility of data from national curriculum tests for between-school comparisons that serve benchmarking and monitoring purposes. We also recognise the utility of basing such comparisons upon aggregated scaled scores, for end-of-key-stage

cohorts. As mentioned above, we believe that there is nothing useful to be gained from reporting comparisons in terms of deciles, which represents a loss of information for no obvious benefit, but with potential costs. Indeed, for benchmarking and monitoring performance, we see no value in reporting other than in terms of aggregated scaled scores. This helps to avoid perverse incentives associated with high stakes cut-scores.

- 4.2. We emphasise, however, that aggregated scores from tests in English and mathematics:
 - i. do not provide trustworthy indicators of the general educational attainment of a cohort (because the majority of intended learning outcomes associated with the full national curriculum are not tested; and some schools may focus their teaching unduly on tested subjects, exacerbating this problem);
 - ii. provide only moderately trustworthy indicators of subject-specific attainment for a cohort (because certain intended learning outcomes associated with mathematics and English, in particular, are not tested; and some schools may focus unduly on tested aspects, exacerbating this problem).

- 4.3. We recommend a reporting strategy of presenting outcomes, and drawing comparative inferences, in terms that are as close as possible to the content tested (e.g. 'reading') rather than in terms of the full subject domain (e.g. 'English') or the full curriculum. We therefore welcome the recent decision to report test results for 'English reading' and 'English grammar, punctuation and spelling' separately from 2013 onwards. This should help to make clear that only the statutory teacher assessments will refer to full subject domains.

- 4.4. To avoid the inevitable temptation for schools to adopt strategies which amount to teaching-to-the-test, it is important to enhance the trustworthiness of data from national curriculum tests through the operation of a strong school inspection regime. The responsibility for ensuring that schools are not unduly restricting their teaching of intended learning outcomes must lie squarely with Ofsted, supported by head teachers and governing bodies.

- 4.5. Evidence of unreliability was central to the change from the original KS1 tests and the shift to statutory teacher-based assessment, which might incorporate some testing but would be balanced by teacher assessment of broader performance. More problematic than unreliability, though, is evidence of negative impact. For example, recent research demonstrates the impacts of date-of-birth – and the long-term effects of these impacts on learning – particularly when younger children are misclassified (in terms of attainment) because there has been insufficient regard to their age. Once again, while data from early years and KS1 tests, alongside other assessment evidence, may be useful for *diagnostic* purposes, their significance

should not be unduly elevated, as would happen if they were to become the basis for school value-added analyses. We therefore recommend that **assessment during reception or at the end of at key stage 1 should not be included within value-added calculations for comparing schools.**

- 4.6. Instead, we believe that primary schools are most appropriately compared using benchmarking techniques that are clearly approximate. This would group schools according to pupil background variables alone (e.g. FSM or equivalent, SEND, EAL), rather than 'prior attainment' which cannot compare like with like because the content of the curriculum changes as pupils move through KS1 and KS2.
- 4.7. As with school accountability measures, Ofsted should be responsible for preparing and reporting these benchmarking analyses, and for synthesising them with additional information on school effectiveness from other available data and from school inspections. The aspiration, here, is for integrative, holistic, fully-informed comparisons between schools, with appropriate caveats.
- 4.8. We welcome the move towards national sample testing at the end of KS2, for the purpose of monitoring national attainment standards over time in science. This approach allows a greater range of learning outcomes to be assessed, thus enhancing the potential for drawing trustworthy inferences from the results. The new tests of English (mainly reading) and mathematics will not involve sample testing; this means that the potential for drawing similarly trustworthy inferences concerning trends in attainment over time will be substantially lower in these subjects. Consequently, it should be recognised that it will not be possible to monitor national attainment standards in English and mathematics, over time with confidence.
- 4.9. Experience also tells us that a substantial, but ultimately unknowable, proportion of the improvement in performance over time – following the introduction or significant revision of high stakes tests – is due to the tests becoming increasingly familiar, and not to genuine improvement in attainment across cohorts. This will exacerbate the challenge of monitoring trends in attainment over time.
- 4.10. In the future, national sample tests in English and mathematics could be developed, in addition to national curriculum tests, to enable national monitoring purposes to be achieved more satisfactorily.

5. Key recommendations

5.1 In summary, our key recommendations are:

- i. Teacher-based assessment and national tests should be identified as separate parts of a national assessment system.
- ii. A national framework for statutory teacher-based assessment should be developed before the new system is introduced. It should subsequently be strengthened by DfE working with schools and teachers, as already proposed in paragraph 2.1 of the consultation document.
- iii. There should be no national collection of teacher-based assessment results; these should only be used for reporting to pupils, parents, receiving teachers and for internal school use.
- iv. Summative reporting should be as informative as possible and avoid the use of arbitrary grades, marks, levels and cut scores.
- v. Reports for individual pupils should provide qualitative accounts of progress in relation to national and/or school curriculum expectations.
- vi. The EYFS Profile for reporting to parents should be retained.
- vii. Any early years/KS1 national tests of phonics/reading should be used diagnostically, not as baselines for national bench-marking and value-added purposes.
- viii. Scaled scores, but not deciles, should be used in reporting national test results.
- ix. The change in terminology to 'English reading' and 'English spelling, punctuation and grammar' should be welcomed as it indicates that only a part of the subject is being tested.
- x. National test results are sufficient for national bench-marking and monitoring purposes at the end of KS2.
- xi. The use of sampling in the monitoring of science standards is welcome; this should be extended to mathematics and English as the best way to monitor standards over time.

6. Conclusion

6.1. The OECD report (2013), *Synergies for Better Learning: an international perspective on evaluation and assessment* (2013), to which we made reference at the beginning of this response, made two further policy recommendations to promote better alignment between policy intentions and practice in education systems.

6.2. First, OECD argue that creating an effective assessment and evaluation system will involve building capacity at all levels of the system. A high quality system will depend on the depth of understanding of those involved in designing, gathering and using evidence. Internationally, there is consistent evidence to suggest the need for professional learning in assessment and evaluation amongst teachers, head teachers and policy makers.

6.3. Consideration should also be given to the challenge of engaging parents in the process, in particular: how to build their understanding of the reporting process; what the evidence tells them about their child; and the limits of its usefulness.

6.4. Second, consistent with wider international research evidence, OECD advise that if an evaluation and assessment framework is to be implemented successfully, a substantial effort should be invested in building consensus among all those involved. People are more likely to accept change if they understand its rationale and potential usefulness. We would urge DfE to consider these further recommendations.

Professor Jo-Anne Baird, University of Oxford
Professor Patricia Broadfoot CBE, University of Bristol
Professor Emeritus Richard Daugherty OBE
Professor John Gardner, University of Stirling
Professor Emerita Caroline Gipps
Professor Emerita Wynne Harlen OBE
Professor Louise Hayward, University of Glasgow
Professor Mary James, University of Cambridge (Convenor)
Professor Paul Newton, Institute of Education London
Professor Emeritus Gordon Stobart, Institute of Education London

1st October 2013