education datalab
research • analysis • evidence

Submission to the House of Commons Education Select Committee Inquiry on Primary Assessment

Dr Rebecca Allen, Education Datalab

Education Datalab is the research arm of the non-profit company FFT Education Ltd who have been supporting schools in their use of data for over 15 years.

Dr Allen is Director of Education Datalab and also an academic on leave from UCL Institute of Education. This submission draws on the work of other members of Education Datalab, notably that of Dave Thomson and Dr Mike Treadaway.

The role of statutory assessment in primary schools

1. Assessment plays numerous roles in primary education. *Statutory* assessment should, as a minimum, ensure **compliance** of schools with the National Curriculum and make schools **accountable** for the quality of the education they provide. This submission provides evidence and a commentary on these dual roles of compliance and accountability.

2. Our **hopes** for statutory primary assessment are that it helps ensure schools teach the curriculum we mean them to teach, that they are incentivised to do it as well as they possibly can, that we can identify, celebrate and replicate outstanding practice and that we can drive out poor quality teaching and ideas.

3. However, we must take seriously and seek to minimise the potential damage that public, statutory assessment can have on children if it unfairly judges good schools, excessively narrows the curriculum, encourages coaching to the test or cheating by teachers, leads headteachers to manipulate their pupil intakes, or creates unnecessary stress for children and teachers.

4. In this submission we provide evidence to make the case that:
    a. Using assessment for curriculum compliance must be done with care
    b. We need a reliable age 5 baseline assessment
    c. Teacher assessment should not be used in statutory assessment
    d. A simpler floor standard would be fairer and more transparent
    e. We should avoid over-interpreting Key Stage data

Using assessment for curriculum compliance must be done with care

5. Tests can only ever measure a sub-set of the wider goals of education. Primary assessment is quite different to Key Stage Four assessment in secondary school because we choose not to assess quite large parts of the primary curriculum. For example, at Key Stage Two the National Curriculum includes subjects such as languages, science, history and music, none of which are tested. We must recognise that what we choose to test alters how what is *not* tested is taught. Many teachers agree that they devote fewer hours to teaching science now that it is no longer tested. Whether or not this matters should be a matter for discussion and enquiry.

6. There is a complex relationship between what is tested and what is taught. One response to a concern that sufficient time is not devoted to a subject is to include it in the test. But, once we reach the stage where most subjects are tested, the consequence is that no teaching time remains for the small areas that are not. Another approach is to *reduce* subjects that are tested, thus reducing overall distortions to curriculum time for other subjects.

7. Curriculum compliance tests such as a phonics check at the end of year 1 can be useful ways of signalling to schools what should be taught and at what stage. There is wide agreement that it has led to more consistent delivery of systematic synthetic phonics schemes in years R and 1.[1] There should now be greater consideration of whether:

    a. The introduction of phonics schemes as early as October in year R disadvantages any pupil groups.

    b. The repeat check at the end of year 2 is sufficient to ensure that those who have not yet learnt to read reach a good standard during primary school.

    c. The intensive focus on phonics in years R and 1 has reduced teaching time in some other areas, such as maths, and whether this matters.

We need a reliable age 5 baseline assessment

8. If we are to measure pupil progress in primary schools then we must stop using the Key Stage One test for this purpose and must instead introduce a reliable and unbiased Reception baseline test that judges children on entry to school. Once we have done this, the Key Stage One assessments will form no clear statutory role.

9. We should not use a test part-way through primary school to baseline attainment because it does not allow us to capture the strengths and weakness of instruction in the lower part of a primary school.

10. The use of Key Stage One assessments as a baseline in accountability metrics provides strong incentives for primary schools to depress their scores. In past research[2] we showed how the replacement of the Key Stage One externally marked test with teacher assessment in 2003 led to primary schools depressing their scores, knowing it would be used as a baseline for Key Stage Two value-added measures. This did not happen in

[1] https://www.nfer.ac.uk/publications/yopc03
[2] http://educationdatalab.org.uk/2015/03/we-worry-about-teachers-inflating-results-we-should-worry-more-about-depression-of-baseline-assessments/

infant schools where Key Stage One is the outcome metric. We could observe exactly the same phenomena during reorganisations from split infant-junior schools into all-through primary schools where the high value-added scores achieved by infant schools from Foundation Stage Profile to Key Stage One would slip away to zero once they became part of a primary school.

11. The pilot of three separate baseline test providers was an admirable experiment in measuring an aspect of test validity – the consistency of alternative measures across the same construct. However, at the time the Department for Education received strong feedback from the profession that using three baseline test providers would not work and, given that the specifications allowed considerable discretion over delivery, it was no surprise that their products were not sufficiently well correlated to be used.[3]

12. That said, this evaluation of the correlates between test providers does not undermine the principle that a carefully designed age 5 baseline (indeed, perhaps one already created by a company) would be more fit-for-purpose than the status quo of teacher assessments at the end of Reception and at age 7.

Teacher assessment should not be used in statutory assessment

13. We currently use teacher assessment in our outcome measures and to create baselines for pupil progress metrics, i.e. in the entire foundation stage profile, across Key Stage One and in the Key Stage Two writing assessment. We should use it for neither baseline nor outcome because it is impossible for someone making a judgement not to be influenced by their knowledge of the use of which the judgement will be put.

14. Teacher assessment should not be used in high stakes settings where **incentives to distort** results are high. As we explain above using the example of the Key Stage One assessment, this is as true in its depression as a baseline as it is as an inflated outcome metric.

15. Teacher assessment is **biased**. Teachers are unable to retain accurate information on dozens of children at once, so they use cognitive shortcuts to rely on stereotypes of performance across groups. For example, if black boys in a class tend to perform poorly in literacy, a teacher is more likely to under-rate the performance of any individual black boy compared to their true capabilities in literacy. This has been shown across numerous academic studies.[4] The result is that teacher assessment discriminates against poorer pupils and minorities which, on its own, is a good enough reason not to use it.

16. Teacher assessment is **unreliable**, particularly where marking criteria not clear. And most strategies to increase the reliability of teacher assessment considerably increase

---

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/514581/Reception_baseline_comparability_study.pdf

[4] See, for example: Burgess, Simon, and Ellen Greaves. Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics* 31.3 (2013): 535-576.
Campbell, Tammy (2015) Stereotyped at seven: Biases in teachers' judgements of pupils' ability and attainment, *Journal of Social Policy*. 44(3): 517-547, doi: 10.1017/S0047279415000227.

workload for teachers and simply make the assessment more test-like. Our research[5] into the 2016 Key Stage Two writing moderation highlighted considerable inconsistencies across local authorities that were not present to the same extent in 2015:

a. The correlation between the percentage achieving expected level in reading and writing across local authorities in 2016 is just 0.35. By contrast, the correlation for reading and grammar is 0.72 and between reading and maths is 0.70. For comparison, in 2015 the correlation between local authority reading and writing achievement was 0.81.

b. The correlation between the local authority percentage achieving expected level in writing in 2016 and achieving a Level 4 in 2015 is just 0.41. The equivalent correlation for reading is 0.85.

17. There is little evidence that training in moderation will sufficiently improve consistency in writing assessments. But there is good evidence that the reliability of the writing assessment could be improved considerably through the use of comparative judgement.[6] This is a process of assessing work through quick, pair-wise comparisons that is designed to collect multiple judgements on every piece, without the need for chief examiners or specific marking criteria.

18. Comparative judgement is particularly well-suited to assessments such as writing where devising a set of criteria against which to mark is almost impossible. The removal of 'tick-lists' of standards reduces some of the less desirable side-effects of teaching to the test. It also works well where the task is open-ended and we want to judge situations where some pupils attempt more challenging work than others. Once in place, it would allow a reduction of some of the more distortionary guidance on punctuation and grammar, such as on 'exclamation sentences'[7], to be replaced by more holistic views of the standard of punctuation and grammar in writing.

A simpler floor standard would be fairer and more transparent

19. Schools should be held accountable for their performance using metrics that are transparent and fair. Whilst secondary schools now know they are subject to a single accountability metric (Progress 8), primary school accountability is based on three value-added measures in reading, writing and maths alongside the percentage of pupils achieving the expected standard in all three subjects. This is confusing to the wider public who want to know whether schools are doing a good job.

20. There are often good reasons for using multiple accountability indicators, for example to represent the complex and multi-dimensional nature of school performance and to discourage gaming. But since the individual subject indicators exhibit much year-on-year variability, aggregating to a single overall indicator can reduce this variability somewhat.

[5] http://educationdatalab.org.uk/2016/09/consistency-in-key-stage-2-writing-across-local-authorities-appears-to-be-poor/
[6] http://educationdatalab.org.uk/2016/10/will-more-training-in-moderation-of-teacher-assessment-help/
[7] This is one of many newspaper items on the issue: http://www.telegraph.co.uk/education/12185164/Nonsense-Backlash-over-new-school-rules-on-exclamation-marks.html

21. We have tested the properties of an overall value-added indicator constructed from performance in maths, reading, and grammar, punctuation and spelling (*not* writing since it is not reliable).[8] Using a metric such as this, with an age 5 baseline, would:

    a. Remove the distortions of the Key Stage One baseline described above.

    b. Remove the threshold expected standard metric from the floor standard, which encourages excessive focus on students at the margin of meeting the standard.

    c. Ensure that all primary schools, even those with very able intakes, are held accountable for poor progress.

    d. Create a more stable and easily intelligible indicator of primary school performance.

We should avoid over-interpreting Key Stage data

22. If we are to use Key Stage tests as accountability devices then we must judge quality fairly and intervene only where appropriate. We do not yet know what the reliability of the new Key Stage Two tests is[9] but, unless we are prepared to spend longer and more money testing pupils, we must recognise they were not designed to measure pupil attainment precisely and can only provide an imprecise picture of school quality.

23. One problem is that a single test needs to measure attainment over such a wide range of pupils. There remain specific difficulties here. Around 3% of pupils were not entered for each of the reading, GPS and maths tests in 2016 because they were working below the standard of the test. Another 1% of pupils took the reading test but were not awarded enough marks to be awarded a scaled score. At the other extreme, just 1% of pupils achieved test marks of 44 to 50 implying that 6 marks of the test were wasted due to being too difficult. An easier test could have provided better measurement of pupils at the lower end of the ability distribution.

24. We cannot get round the fact that primary schools are relatively small and we usually try to judge school performance on just 30 to 60 pupils (and sometimes even fewer). This is one reason why the year-on-year correlation in school test performances is less than 0.6,[10] yet our best guess is that true variation in school practices is not so volatile. We use confidence intervals on school performance metrics to reflect this sampling variability, but they do not account for other aspects of the unreliability of measurement.[11]

---

[8] http://educationdatalab.org.uk/2016/10/should-ks2-floor-standards-be-based-on-an-overall-value-added-score/

[9] Ofqual studied Key Stage Two reliability as part of a wide-ranging study on the topic: https://www.gov.uk/government/collections/reliability-of-assessment-compendium

[10] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/549432/SFR39_2016_text.pdf

[11] http://educationdatalab.org.uk/2015/07/significance-tests-for-school-performance-indicators-why-theyre-ok-really-but-we-should-probably-call-them-something-else/

25. In summary, there are multiple reasons why the judgements we make on schools are so fragile:[12]

    a. Tests are not perfectly reliable and create a band of uncertainty around a pupil's test score

    b. Pupil test scores are affected by the choices of content, item formats, scoring methods and scaling decisions

    c. Using aggregated pupil test scores to infer school performance introduces sampling error

    d. Our choice of aggregated metrics will change the rank order of schools, in part because so many schools differ so little in their practice and performance.

    e. Intense pressure to raise scores creates the potential for distorted metrics of true learning gains and we do not know how much schools vary in their distortions.

26. We therefore believe we should *lower the stakes* associated with a single year of poor primary school performance. Deviations from expected performance should simply be *worthy of investigation* rather than deserving of definitive judgement. In the case of primary schools which tend to be quite small, stronger judgements should only be made on multiple years of school performance data.

27. Lowering the stakes associated with one year of poor performance is not only fair, it also has a potential to reduce some of the deleterious effects associated with statutory assessment.

[12] This is a summary of the points made by Daniel Koretz in his book Measuring Up.