Assessing achievement at the end of key stage 2

Dylan Wiliam
Emeritus Professor of Educational Assessment
Institute of Education, University of London

Introduction

There is solid, and growing, evidence that the presence of high-stakes accountability systems, in which local authorities, schools, teachers, and, ideally, pupils are held accountable for their academic performance, is associated with increased pupil achievement. A recent review of the research in this area (Wiliam, 2010) concluded that the effect of such assessment regimes was equivalent to an increase in the rate of learning of approximately 20%. High-stakes accountability systems are probably the most cost-effective method for raising achievement yet developed.

However, in *every single* instance in which high-stakes accountability systems have been implemented, adverse unintended consequences have significantly reduced, and in many cases have completely negated, the positive benefits of such regimes. The question is, therefore, can we do better than the architects of these existing systems, and secure the positive benefits of high-stakes accountability testing, while at the same time, minimizing the unintended adverse effects? To answer this question, it is necessary to understand how the unintended effects arise.

Unintended consequences of high-stakes testing

The most commonly cited criticism of high-stakes testing is that it results in "teaching to the test". However, it is important to realize that this is an adverse consequence only if the test does not adequately cover all the intended learning outcomes. If the test assesses all the valued outcomes of learning, then teaching to the test is exactly what we want teachers to do. The difficulty is that we cannot, in the time that is likely to be made available for testing, cover everything that has been taught, so it is necessary to sample. Where the sample is random, then teachers cannot "guess" which topics will come up, and so they have teach the whole curriculum—we have, in the words of Lauren Resnick, "tests worth teaching to". Where the sample is not random, and it is possible to predict what will, and will not, be tested, and the stakes are high enough, this creates an incentive for teachers and pupils to put more emphasis on those elements of the curriculum that will be tested. In the past, successive governments have assumed that such adverse effects do not exist, are not harmful, or can be minimized through appeals to the professionalism of those being held accountable. The evidence from all over the world on the impact of testing regimes is that none of these assumptions is reasonable.

At the end of key stage 2, in particular, we have seen a progressive focusing on teaching pupils what will be tested. From the early days of national curriculum assessment, the curriculum for year 6 in many primary schools gave greater weight to English, mathematics and science than to the other, statutorily required, but untested, national

curriculum subjects. Within each of the tested subjects, the first attainment target in each subject (Speaking and listening, Using and applying mathematics, and Scientific enquiry) was explicitly excluded from the test, and as a result, less attention was given to these aspects of the curriculum. Even within the attainment targets that were explicitly addressed by the tests, teachers can predict which aspects can, and cannot be addressed given the format of the test, and spend more time on those that will be tested.

The consequences of these unintended consequences are profound. As well as making learning less interesting for pupils, the test results become increasingly useless as indications of pupil achievement. At the national level, as the work of Peter Tymms has shown, increases in the proportion of pupils reaching level 4 in English and mathematics are not matched by increases in scores on standardized reading and mathematics tests. Our pupils have got better and better at less and less. At the level of the individual pupil, secondary schools have found that the achievement at key stage 2 becomes almost useless as an indication of what pupils have learned, and how the school can best meet their needs.

There is, in fact, ample evidence that teaching to the test is *not* the best way to increase test scores but it appears that it is very hard, if not impossible, to get most teachers to make the "leap of faith" necessary to resist narrowing the curriculum. The best working assumption is that if pupils' score can be increased by ignoring some part of the curriculum, it is likely to happen to a significant extent. The only way to ameliorate this situation is to broaden the basis on which pupils are assessed, and this requires increasing the amount of time spent assessing.

Increasing the length or number of the tests at the end of key stage 2 will be expensive, unpopular, and difficult to justify. In the words of Peter Silcock, "schools are places where learners should be learning more often than they are being selected, screened or tested in order to check up on their teachers. The latter are important; the former are why schools exist." For that reason, there has been significant interest in making greater use of the information about pupil achievement collected by teachers as a normal part of their work.

Teacher assessment

Many of the claims about the superiority of teacher assessment over traditional tests are, quite simply, fanciful. Since the volume of the work being assessed is greater, the reliability of the assessment is likely to be higher, but there is strong evidence that teachers take a number of irrelevant factors into account in assessing pupil work. Many studies have shown that even when they are meant to be assessing achievement, teachers are swayed by pupils' behaviour, attitude, and even facial attractiveness. Most recently, the work of Paul Black and others at King's College London has shown that even specialist mathematics and science teachers in the secondary school cannot be relied on to generate valid assessments of their pupils' achievement.

The most common method for addressing the difficulty that teachers have in generating fair and objective assessments of their pupils is group moderation. The idea here is that teachers assess samples of their pupils' work, and compare them with those of other

teachers. However, the small amount of evidence about these moderation meetings that does exist suggests that while such meetings can be a valuable source of professional development, they are not robust mechanisms for aligning teachers' judgments in the short to medium term. What we need is a way of combining the knowledge that teachers have about their pupils with hard-edged information that allows us to ensure that consistent standards are being used in every primary school in England.

A proposal for national curriculum assessment at key stage 2

During the year, each teacher would collect whatever records on pupil achievement that were most useful for her teaching. The records would be integrated into the curriculum scheme so that the record was genuinely useful for planning (see Clymer & Wiliam, 2006/2007 for an example of how this could work in practice).

At some point in the year, one day would be designated for formal external assessment (half a day for English and half a day for mathematics). Ideally this would be towards the end of the school year in order to capture the 'latest and best' achievements of pupils but could be conducted as early as March. Prior to the testing week, the school would submit to the DfE the names of all pupils in year 6. From this list of names, pupils would be randomly allocated to receive particular assessments. For example, in a single-form-entry school the allocation for the half-day for English testing might be as follows:

4 pupils are given a reading test
4 pupils are given a written spelling test
4 pupils are given a handwriting assignment
4 pupils work on a collaborative assignment to assess speaking and listening
4 pupils are given a creative writing assignment
4 pupils are given a factual writing assignment
4 pupils are given an assignment on writing in another national curriculum subject.

Some of these assessments could be marked by machines (with the increasing availability of automated scoring of natural language, both multiple-choice and constructed-response items could be scored automatically). Others would be marked by hand, as is the case currently.

The score achieved by a pupil on a particular task would not be a particularly reliable indicator of her or his achievement, nor would the score of the four pupils allocated to the handwriting assignment necessarily be a good guide to the level of handwriting in the class. However, the average achievement of the class on *all* these tasks would be a highly reliable indicator of the *distribution* of achievement in the class. The results of the assessments would be reported to the teacher as a profile of levels. For example, a teacher might be informed that in her class, in English, six pupils were assessed as level 3, twelve as level 4 and ten as level 5 but the teacher would not be told which pupils had scored at which level on the tasks. The teacher would then have to a level for each pupil within the envelope of levels given, by using the records of the pupils' achievements that she had been maintaining throughout the year.

Because the teacher's assessment would be based on hundreds of hours of assessed work by the pupil, the resulting assessments would be highly reliable, and because the levels awarded are constrained to fit the envelope of levels defined by the external tests, they would be comparable from school to school. Moreover, since the teacher does not know which pupil will be assessed on which basis in the external tests, there is no incentive for the teacher to 'teach to the test'. Or, more accurately, the only way to 'teach to the test' is to improve the capability of *all* the pupils on *all* the tasks, which is surely what we want.

An additional benefit of such a matrix-sampling design is that it allows new kinds of assessments to be introduced over time. In the initial years, it would probably be necessary to start with fairly conventional assessments. When new forms of tasks, such as co-operative group tasks are to be introduced, teachers could be given plenty notice of the change, and appropriate professional development support could be made available.

The main weakness of the model is that it is not direct, and this produces three specific difficulties. First, there is a lack of transparency, in that it is not possible to count up the marks gained by a pupil on the assessment to determine the level that the pupil receives. Such lack of transparency may be a source of confusion, but the challenge of communicating this to stakeholders does not seem insurmountable.

Second, allowing teachers to determine which pupils are given which levels may allow the teacher's personal bias to influence the levels given. This is undoubtedly a risk, but can be monitored by looking at the correlation between levels awarded and scores obtained on the assessments. If the pupils getting high scores on the tests were not being given the highest levels, then this would be obvious, and could be investigated.

Third, and perhaps most significant, the model depends on pupils being motivated to do well, even though the impact of their performance on their own score is small. It would therefore be important for the school to ensure that all pupils felt it was important for them to do well for the benefit of the whole school—no bad thing in itself.

The broad framework outlined above could be implement in a number of ways. None of them will be perfect, and all will present significant challenges. Nevertheless, the 'big idea' presented here—sampling of pupils to define an envelope of levels within which teachers exercise judgment in awarding levels to pupils—appears to have significant potential for avoiding the narrowing of the curriculum caused by timed written tests, while at the same time providing levels of achievement that are comparable from school to school and authority to authority.

References

Clymer, J. B., & Wiliam, D. (2006/2007). Improving the way we grade science. *Educational Leadership,* **64**(4), 36-42.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist,* **45**(2), 107-122.