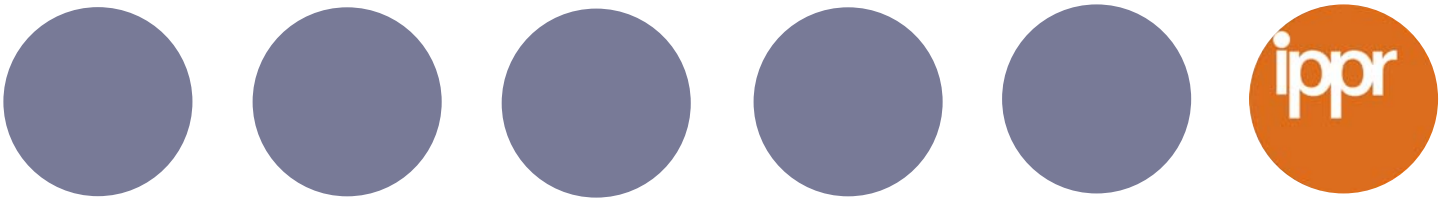

Assessment and Testing:

Making space for teaching and learning

Richard Brooks and Sarah Tough

DECEMBER 2006

Institute for Public Policy Research
www.ippr.org

The Institute for Public Policy Research (ippr) is the UK's leading progressive think tank and was established in 1988. Its role is to bridge the political divide between the social democratic and liberal traditions, the intellectual divide between academia and the policymaking establishment and the cultural divide between government and civil society. It is first and foremost a research institute, aiming to provide innovative and credible policy solutions. Its work, the questions its research poses, and the methods it uses are driven by the belief that the journey to a good society is one that places social justice, democratic participation, economic and environmental sustainability at its core.

This paper was first published in December 2006.

30-32 Southampton Street, London WC2E 7RA
Tel: 020 7470 6100 Fax: 020 7470 6111 www.ippr.org
Registered Charity No. 800065

Acknowledgements

This work has been made possible through the generous support of Cambridge Assessment and Select Education Plc. The authors would like to thank all those who have contributed to the ideas contained in this paper, through seminars and informal discussions, especially Professor Dylan Wiliam from the Institute of Education.

In particular we would like to thank Sylvia Green, Bene't Steinberg and Tim Oates from Cambridge Assessment for their expert contributions. Cambridge Assessment is publishing their own independent work on assessment in schools which is available at www.cambridgeassessment.org.uk

We would further thank Jodie Reed and Peter Robinson for their work at ippr during the early stages of the project as well as Richard Darlington and Nick Pearce for their continuing advice and support.

This paper is part of a broader project on school reform. It is published alongside a companion paper on pupil attainment, and further papers on school admissions and funding are due to be published in 2007.

The views expressed in this paper are those of the authors alone.

About the authors

Richard Brooks is an Associate Director at ippr and leads the institute's public services work. He is a frequent contributor in both the broadcast and print media, and has published widely on public services and welfare policy issues. Prior to his current role he was Research Director at the Fabian Society, and he has also worked in the Prime Minister's Strategy Unit and the Labour Party Policy Unit. Between 2002 and 2006 he was a Tower Hamlets Councillor, and was Cabinet Member for Resources for three of these years. He is a governor of Bethnal Green Technology College.

Sarah Tough is a Research Assistant in the Public Services Team at ippr. Before joining ippr Sarah worked as a mathematics teacher in a London secondary school as part of the Teach First programme. Sarah has a first class honours degree in Economics from the University of Bristol. She is a governor of a Hackney primary school.

Overview

Curriculum, pedagogy and assessment define what is taught, how it is taught, and how we know that learning has taken place. Together they describe much of what takes place in the classroom, and the relationship between them is crucial. This paper focuses principally on assessment, which has increased in importance over the past ten years as a tool of schools policy. Assessment, and in particular testing, now defines much of what goes on in schools, from decisions around resources to teaching strategies in the classroom.

If we are going to hold schools to account on the basis of assessment data, such as national curriculum test and GCSE results, then what is assessed and how will have profound consequences for what is taught and how. This is of course the very intention of the current system. However, we will argue here that the current assessment system is having unintended and unwelcome consequences for the quality of teaching and learning.

This paper is primarily concerned with assessment in the form of national testing up to the end of Key Stage 3 at age 14. We do not examine the accreditation and qualification framework.

Objectives for the assessment system

Pupils, teachers and parents all need assessment for different reasons. Pupils need assessment to assist them in their learning. Teachers need assessment so they can understand their pupils' needs and consequently adapt their teaching. Parents need assessment so that they can understand how their children are developing and how well their school is performing. Head teachers, governors, local authorities and central government all need assessment to hold schools accountable for the outcomes they achieve.

More formally, the assessment system as a whole should achieve the following aims:

1) It should be valid and reliable. The assessment system should measure what it claims to measure – generally, the degree to which a pupil has learned and understood a subject or skill. In addition, it should produce reliable results: the measurement should be reasonably accurate and not suffer from a lot of random variation between schools or over time.[1]

2) It should periodically identify and record the level of achievement that individual pupils have reached. Such certification is probably the most traditional function of the assessment system, and is often called 'summative' assessment.

3) It should allow parents and others to hold learning institutions accountable for their performance. This is probably the aspect of assessment that has most increased in importance over the last decade with the arrival of school performance tables and a system of national targets for key stage results. This function has previously been described as evaluative assessment, where it is the education service rather than the pupil that is being evaluated.

4) It should facilitate the learning of the pupil. Assessment should help identify the gaps and weaknesses in an individual's understanding, as well as the strengths that can be built upon, so as to help inform the teaching and support subsequently received by the pupil. Such assessment is often now called formative assessment, or assessment for learning.

5) It should be clearly understood and enjoy public trust and confidence, particularly among parents and teachers.

6) It should enable the monitoring of national standards over time.

It is important to note that these are objectives for the assessment system as a whole, not for each component part. One of the key arguments we make later in this paper is that different forms of assessment are better suited to achieving different objectives, and that our current emphasis on one form of assessment – the national curriculum tests at the end of the Key Stages – is having undesirable side-effects.

1. For a detailed discussion of reliability and validity see Wiliam (2000a, 2000b).

How much weight should we accord each of these objectives? To start with, validity and reliability are a prerequisite of achieving the others. Whatever use assessment is put to, it must measure with reasonable accuracy what we want it to measure. As we will see later in this paper, this is more demanding than it sounds. Trust and confidence, meanwhile, should be the outcome of a system of assessment that is widely seen to meet the other criteria effectively.

When and why do we need summative assessment? During the course of schooling, parents want to know that their children are making sufficient progress, and often pupils themselves do too. Summative assessment also helps pupils and parents choose options such as GCSE subjects or a specialist secondary school. However, the major occasion for such assessment occurs at the end of secondary school, when GCSEs and/or GNVQs (General National Vocational Qualifications) become the passport to further education and employment. This objective is thus important, but much more important at the end of schooling than before this point.

Since the 1970s schools and teachers have become much more accountable for what they do and for the outcomes achieved by their students. This is absolutely right and appropriate. Accountability means that parents and the public can understand what is happening in each school (and in the schools system as a whole); that schools take responsibility when things go wrong (and also when things go well); and that schools are responsive to pressure for change and improvement. All of this requires that good-quality information on school performance is available, and assessment is one of the critical means of providing this. If we are going to recommend changes to the assessment system, we need to be very careful that they do not weaken school accountability.

Perhaps the most important role of assessment is to help pupils learn. Assessment for learning is sometimes described as 'the new pedagogy', but it has been a live issue in education theory for some time (see Box 1).

Box 1 Assessment for learning

Assessment for learning, or formative assessment, has existed as a distinct idea for many years (see for example Bloom *et al* 1971), and refers both to a body of theory and to a range of classroom practices. Its objective is to use pupil assessment as a bridge between current performance and desired achievement, by generating information about each pupil's development and then using this to inform their teaching appropriately. Assessment for learning is thus essentially cyclical, with the information gained by teachers and pupils in each cycle of assessment being used to adjust teaching and improve performance the next time around.

While assessment for learning is not just a set of classroom practices, it is associated with some specific techniques, including, for example:

- Starting lessons with a question, and explicitly sharing the objective of the lesson with the pupils
- More open questioning by teachers, with longer waiting times between questions and answers, allowing pupils more space in which to think
- 'No hands up' questioning, where everybody is expected to be able to offer an answer, and 'everyone answers' questioning where all pupils record an answer and hand it in
- Comments-only marking, indicating strengths and how to improve but without providing scores to distract pupils
- Self assessment by pupils, for example where they indicate via a traffic light system whether they feel they have made the necessary progress
- Peer assessment and teaching, where pupils help each other with their learning.

There is now an extensive body of evidence indicating the effectiveness of assessment for learning as a pedagogical tool. Black and Wiliam (1998a) provide a wide-ranging survey of 250 studies that indicates that effective use of classroom formative assessment – with a short cycle of assessment, feedback and changes to teaching – approximately doubles a pupil's rate of progress across a wide range of subjects and contexts. In addition, it does not depend on a radical increase in school inputs such as staff, IT equipment or buildings. Thus in one sense it is a way of increasing teachers' productivity. It may, however, require significant training and represent a major challenge to some teachers who are required to change their most basic classroom practices. Many of the teachers involved in assessment for learning experiments participated in multiple half- or whole-day training workshops over a period of more than a year.

The most effective schools now practise a culture of continuous teacher-led assessment, appraisal and adjustment of teaching practices to personalise learning for all their pupils. It seems clear that assessment that does not assist the learning of the child is of very limited value, and in many ways the certification of achievement and the accountability role of assessment are only important because of their links to this.

The public debate about assessment in schools often seems self-contradictory. Those who care about equality sometimes call for an increased focus on low-attaining pupils at the same time as complaining about the burden of assessment – presumably over concern about the means of identifying the relevant pupils in the first place. Meanwhile, if national test results go up, some will take this as evidence that the tests are getting easier, while failure to make progress towards the relevant target will also be criticised.

As we will go on to explain, these apparent contradictions can be addressed through changing the nature of the assessment system. We are not simply thinking in terms of 'more' or 'less' assessment, but of changing the nature of assessment mechanisms and the shifting balance between them.

How does the assessment system currently work?

The foundations of the current system of assessment were brought into force by the Education Act of 1988. National testing at the ages of seven, 11, 14 and 16 accompanied the introduction of the National Curriculum, which for the first time specified programmes of study and attainment objectives for all pupils attending maintained schools. There were many benefits to the introduction of the National Curriculum, in particular improved coherence across hitherto uncoordinated geographical areas and different phases of schooling. Training for teachers also accompanied its roll-out. The system of a National Curriculum, with national tests and target levels of attainment at the end of each key stage, is still in place today.

The initial roll-out of national assessment was accompanied by a move to capture the potential of assessment as a pedagogical tool, in line with the intentions of the independent Task Group on Assessment and Testing (TGAT) that was set up to advise the government of the day on these matters. Yet the Conservative government's emphasis on using parental choice of school as an incentive for schools to improve their performance, and the accompanying stress on results as a very public form of school and teacher accountability, led the pendulum to swing away from formative uses of assessment in England over the 1990s[2].

The model that became dominant is sometimes described as a 'high-stakes' assessment system. The stakes were high for schools first because their results were published for parents and the public to see, and second because poor results would attract the attention of the schools inspectorate, Ofsted. One important concern is that such high stakes systems may give schools perverse incentives to do things that are not in the best interests of their pupils, an issue we investigate below. However, we do want to give schools strong incentives of the right kind to help their pupils do well, and we certainly want to be able to identify under-performing schools so that we can intervene where necessary. To some degree there will thus always be a 'high stakes' aspect to an assessment system that holds schools accountable.

Labour governments since 1997 have broadly endorsed, developed and placed more weight on the system of national tests taken in key subjects by all children at the end of Key Stages 1 to 3. Great emphasis has been placed on increasing the proportion of pupils who achieve the target levels in these tests, and on minimising the number of schools where unacceptably low proportions of pupils do so. This is the standards agenda; we discuss the outcomes in terms of attainment in another paper (Brooks and Tough 2006), where we emphasise the need for standards to continue to improve, and for the attainment gaps between different groups to close.

A sustained faith by the Government in the possibility of a quasi-market in school places, in which parental choice is meant to drive up standards[3], has meant a continued emphasis on school performance tables

2. In 1988 the Government accepted proposals for national assessment put forward by the Task Group on Assessment and Testing that included assessment at seven, 11 and 14, driven by an aim to support formative practice. It concluded that there should be a framework of clear criteria, or criterion referencing, spanning age ranges, so that teachers could make best use of assessment to inform learning. Under the TGAT's proposals, teacher assessment was combined with external tests and standards were to be maintained by teachers comparing results with the results of the national tests and with the judgments of other teachers. It emphasised that league tables should not be published. For a historical account of how these proposals were gradually abandoned, see Daugherty (1995).

3. We will address school admissions policy in a separate paper.

('league tables'), including their extension to Key Stage 3. New, 'value added' measures of attainment that focus on pupil progress have been introduced to give a truer picture of school performance than that provided by raw results, and from 2007 'contextual value added' data will take into account the individual and family characteristics of pupils to further improve the measure of school performance. Results are now used to inform school decisions about performance-related pay, to inform Ofsted decisions about whether schools should be given light or heavy touch inspections and, combined with targets, to inform judgments about the efficacy of educational initiatives such as the Primary Strategies.

The consequence of using pupil assessment to judge teachers and institutions in this way has been that external testing has decisively eclipsed teacher assessment for all phases beyond Key Stage 1, where the mode for national assessment was reverted to placing a much greater emphasis on teacher judgments with only a Teacher Assessment level being reported nationally from September 2004[4]. Pupils still sit tests at Key Stage 1, but these are used to corroborate and inform the teacher's judgment alongside references to the pupil's work over the year. There is, however, considerably more flexibility in terms of which tasks/ tests they can use and when (see www.teachernet.gov.uk/educationoverview/briefing/news/ks1assessment/).

Yet at the same time the Government has recognised the need to both make better use of teaching professionals, and to deliver more personalised teaching, through assessment for learning. This can be thought of as operating at two new and distinct levels.

The first level is a more nuanced use of data in school and system management. Individual, pupil-level national assessment data, initially collated for the purposes of measuring the value added by schools (and hence holding schools accountable), has been built on and assimilated into sophisticated data banks that provide a highly versatile tool for a sensitive and contextualised version of performance management – or what former Schools Minister David Miliband has termed 'intelligent accountability' (Miliband 2003). Cross-system data, and school-level data, aggregated from individual performance measures can be used for monitoring performance issues, evaluating progress, diagnosing difficulties, setting appropriate targets based on a full understanding of what might be achievable, and deciding where to channel effort. The schools system has become incredibly data-rich.

On one hand this has facilitated top-down performance management by central government. However, the data has also been made accessible at the local level, transforming it into a potentially highly valuable diagnostic tool. Through the allowing of local access to contextualised data via Ofsted's Performance and Assessment (PANDA) Reports, the Department for Education and Skills (DfES)'s Pupil Achievement Tracker software (PAT) and its successor RAISEonline (Reporting and Analysis for Improvement through School self-Evaluation), there is now potential for local authorities and school leaders to use the data to set appropriate attainment targets for their pupils, to assess their progress against that of similar pupils elsewhere, and to compare their own test results against good national comparators. The level of detail available goes down to being able to compare different groups of pupils' success with different types of questions. Assessment has become a powerful tool for supporting informed and rigorous self-management and we should be careful not to lose this valuable information in any reform of the system.

The second level is the positive promotion of formative assessment as one of the most effective pedagogical approaches. Although originally underplayed by the Government in the wake of TGAT, research has convincingly shown that formative assessment is one of the most effective pedagogical approaches for improving attainment, differentiating teaching and nurturing vital skills such as self-regulation (Black and Wiliam 1998, 1998a). As a result, the idea of using frequent, interactive teacher- and pupil-led assessment is being widely endorsed and is now viewed by many as the ultimate tool for personalised learning.

DfES and associated governmental agencies such as Ofsted have produced a wealth of advice and materials for teachers specifically aimed at increasing understanding of effective formative assessment practices, for example adopting the Assessment Reform Group's '10 Principles for Assessment for Learning' and adapting it into a poster for schools. Meanwhile, the Qualifications and Curriculum Authority (QCA) has been asked to review the Key Stage 3 curriculum, specifically with a view to giving advice on strengthening teacher assessment in foundation subjects, and to developing a bank of tasks relating to attainment targets that can form the basis for formative assessment (although external assessment remains a non-negotiable staple for end-of-phase assessment in core subjects).

4. The first annual statistics reporting teacher assessment only therefore come from the 2005 data.

While proponents of formative assessment have traditionally placed emphasis on comment-only marking (to avoid pupils focusing only on their marks), software drawing on pupil-level data creates the possibility for teachers to use performance data as a starting point for formulating rigorous and appropriate targets for individuals that take into account realistic projections of how similar learners have progressed. Individual question analysis can be used to explore pupil strengths and weaknesses in particular modes of thinking as well as areas of study.

In summary, the current system continues to place enormous weight on national tests at the end of the key stages, while showing signs of a newer emphasis on assessment for learning. The big question is: do the various elements of the assessment system fit together in a way that successfully achieves our objectives? How does the current system match up to our objectives of validity and reliability, providing appropriate measures of achievement, ensuring accountability, facilitating learning, and achieving public understanding, confidence and trust? This is explored in more detail below.

Does the current system meet our objectives?

Validity, reliability and recording achievement

Unfortunately, it turns out that the existing key stage tests are not very reliable at correctly classifying the level of attainment of individual students. The fundamental reason for this is that they rely on a series of tests that can only cover either a small area of the curriculum in detail, or a broad area very lightly. Work undertaken by Dylan Wiliam has estimated, with generous assumptions about the reliability of the key stage tests, that around 32 per cent of Key Stage 2 results and around 43 per cent of Key Stage 3 results are misclassified by at least one level. For a technical discussion of this work see Wiliam (2000b, Black and Wiliam 2006). See Box 2 for a discussion of the reliability of the '11-plus' in Northern Ireland.

Box 2 Northern Ireland and the Transfer Test

The Transfer Test in Northern Ireland is taken at age 11 by pupils in the province who wish to attend a grammar school there. The Transfer Test consists of two exams each worth 75 marks, and each including questions on English, mathematics and science. The scores achieved enable the pupils to be placed in rank order and then allocated a grade (A, B1, B2, C1, C2, D). Grammar schools use the grades to differentiate between pupils applying for places at their school, and only 30 to 40 per cent of those who take the test will be offered a place. These are high stakes tests for those who take them.

An independent empirical study (Gardner and Cowan 2000, 2005) looked at the reliability and validity of the Transfer Test. It concluded that the tests were not even reliable enough to justify the allocation of different grades, let alone the inferences drawn from the grades and the school allocations made on their basis. Due to the narrow range of scores the grades cover, Gardner and Cowan predicted that approximately 70 per cent of those taking the test could be misclassified (ibid). Although the process is complicated by the special arrangements for government in Northern Ireland, it is now expected that the Transfer Test will be abolished by 2009.

One apparently obvious solution for improving the validity of the tests would be to make them longer and thus cover a broader range of material. Unfortunately, the accuracy of the test only creeps up very slowly as its duration is increased, so that increasing the test by a reasonable amount of time will only slightly reduce the numbers of pupils being misclassified. To ensure that Key Stage 2 tests classify no more than 10 per cent of pupils incorrectly, the tests would have to be over 30 hours long for each subject (Black and Wiliam 2006). This is not a problem specific to the current design of the key stage tests; rather it is an inherent problem with examinations where every entrant has to answer the same set of questions to test a reasonably extensive subject, so it applies equally to many qualifications.

A more fundamental problem emerges when we look for evidence of the validity of the key stage tests. These tests are focused on the core subjects of literacy, numeracy and science, on the basis that they represent core skills that are vital to every young person's future development. We thus need to be confident that they are providing a valid measure of pupils' true abilities with respect to these core skills.

In order to consider evidence for or against this hypothesis we would need to examine results from the national tests with a different, independent measure that reasonably claimed to be testing the same skills. If key stage test results mirror the independent measures, for example if both improve over the same time period, then this is some corroborating evidence that both are valid. However, if key stage test results are

going in one direction while the independent measures are going in the other, then this is evidence of a problem with at least one of them.

The question of national standards recurs every time results are published for key stage assessments, GCSEs and A Levels. In summary, we believe that there has been real progress in each of the three core subjects, but less than is indicated by the key stage results. We do not think that the tests have become systematically easier[5]; rather, we believe that teaching and learning has focused more and more narrowly on achieving test results.

Professor Peter Tymms pulled together much of the available evidence on standards in primary schools in his paper 'Are standards rising in English primary schools?' (Tymms 2004). This is the area that has seen the greatest increase in measured national standards, but these improvements have not been sustained at secondary school as the same cohorts of pupils take their Key Stage 3 tests. Using eleven independent measures of English (reading) and mathematics in primary schools over the period 1995 to 2003, Tymms finds that during the period 1995 to 2000 the dramatic increase in the national measure (that is, the Key Stage 2 results) appear to be overstated. The proportion attaining the target level in English at Key Stage 2 rose from 49 to 75 per cent over the period 1995 to 2000, and this equates to two standardised points a year (this is the same in mathematics). The data from the six independent sources and the corresponding Key Stage 3 (matched to the relevant cohort of pupils) results do not corroborate such a striking rise, showing only an average rise of 0.77 points per year for reading and 1.54 for mathematics (Tymms 2004).

The evidence on standards over time is complex. Massey *et al* (2003) investigated whether the difficulty of the tests has changed over time. They found that at least half the national gains in English at Key Stage 2 from 1999 to 2002 were attributable to changes in standards of the tests. A smaller study by Green *et al* (2003) found that there had been improvements in writing over the period 1995 to 2002 and that marking had remained consistent between these years (Green *et al* 2003). While it is possible that there has been some inconsistency in test difficulty, it seems unlikely this is a major or systematic factor.

Teacher assessments have been collated alongside the key stage test results since 1995. Despite the Government officially viewing teacher assessment as 'an essential part of the national curriculum assessment and reporting arrangements' (TeacherNet 2006: 3), interest in teacher assessments and consequently their inclusion in 'league tables' has declined.

Table 1 compares the test results from Key Stages 1, 2 and 3 with the teacher assessments. Key Stage 1 teacher assessment figures closely matched the test results, which is not surprising given the close relationship between the two forms of assessment at this key stage (Richards 2005). However, with later phases teacher assessments provide an interesting contrast to the key stage tests. Key Stage 2 results show an increase in 20 percentage points from 1995 to 2006 in the proportion obtaining a level 4 or higher in English compared to 30 points in the external assessment (see Table 1, next page). Mathematics demonstrates a similar trend. In essence, teacher assessments of pupil performance started slightly higher but have improved more slowly than the national tests. Key Stage 3 teacher assessments replicate the same pattern with tests reporting more improvement than the teacher assessments.

An international survey looking at mathematics and science attainment, 'Trends in International Mathematics and Science Study' (TIMSS), shows some evidence from Year 5 pupils to support raising levels of overall achievement between 1995 and 2003 in mathematics and science. Question items common to both rounds of tests allow comparisons to be made between the two years 1995 and 2003 and in both subjects the percentage answering these questions correctly increased (by 9 percentage points in mathematics and by 4 in science) (Ruddock *et al* 2004). Although the two are not directly comparable, improvements in TIMSS are thus much less impressive than the measured improvements in key stage test results.

The Statistics Commission considered these issues in 2005 and concluded that:

> 'The Commission believes that it has been established that (a) the improvement in Key Stage 2 test scores between 1995 and 2000 substantially overstates the improvement in standards in English primary schools over that period, but (b) there was nevertheless some rise in standards.' (Statistics Commission 2005: 4)

5. There is, however, evidence that standards in some subjects have varied over time. For example, Massey *et al* (2003) found variation in standards in Key Stage 2 English between 1996 and 1999/2000, Key Stage 2 science between 1996 and 2001 and Key Stage 3 mathematics between 1996 and 2001.

Table 1 Pupils achieving the target level in key stage tests and teacher assessment

		Reading		Writing		Mathematics	
Key Stage 1		Test	Teacher assessment	Test	Teacher assessment	Test	Teacher assessment
	1995	78%	79%	80%	77%	79%	78%
	2004	85%	85%	82%	83%	90%	89%
	Point difference	7	6	2	6	11	11
Key Stage 2		English				Mathematics	
		Test	Teacher assessment			Test	Teacher assessment
	1995	49%	57%			45%	54%
	2006	79%	77%			76%	78%
	Point difference	30	20			31	24
Key Stage 3		English				Mathematics	
		Test	Teacher assessment			Test	Teacher assessment
	1995	55%	63%			58%	62%
	2005	74%	71%			74%	75%
	Point difference	19	8			16	13

Notes: 1. The 2006 data is provisional 2. The target or 'expected' level at Key Stage 1 is Level 2, at Key Stage 2 is Level 4 and at Key Stage 3 is Level 5. 3. The data include all eligible pupils in maintained schools and in independent schools that opted to take part in the National Curriculum assessments. 4. For Key Stage 1 2004 figures are used, as the assessment method changed for the 2005 assessments and therefore the figures from 2005 onwards are not directly comparable to those prior to that year. For more details on the new arrangements see Shorrocks-Taylor *et al* (2004). 5. For Key Stage 3 2005 figures are used, as the 2006 figures based on teacher assessments have not yet been published (as at 11 December 2006) due to inconsistencies in the data.

Source: Teacher assessment data (DfES unpublished note 2006), test data see DfES (2006a, 2006b, 2006c)

Looking at the secondary phase, the percentages of pupils attaining the benchmark at Key Stage 3 and Key Stage 4 have continued to rise although progress on international attainment measures has stalled. Evidence from TIMSS for Key Stage 3 (Year 9) does not show any significant change in performance between 1995 and 2003 (Ruddock *et al* 2004). Analysis of the international study PISA (Programme for International Student Assessment)[6] shows that for a given score at Key Stage 3 or Key Stage 4, pupils attained on average a higher PISA score in 2000 than in 2003 (Micklewright and Schnepf 2006)[7]. One

6. PISA looks at attainment in reading, mathematics and science literacy across participating countries (mainly OECD countries) every three years. The analysis described here (Micklewright and Schnepf 2006) compares reading and science literacy in 2000 and 2003. Mathematics is not included as the content areas used for measuring mathematics ability were different between 2000 and 2003.

7. Research by the National Foundation for Educational Research (NFER) for DfES compared the familiarity and appropriateness of PISA and TIMSS for English pupils who take Key Stage 3 tests and GCSE examinations. For science the familiarity suggested that 40-50 per cent of pupils would be familiar with the PISA and TIMSS tests whereas for mathematics TIMSS had a higher familiarity rating of 65-85 per cent compared to 50-70 per cent for PISA (Ruddock *et al* 2006). The focus of PISA is on literacy and this is reflected in the PISA tests, which require much more reading than TIMSS, Key Stage 3 or GCSE. The PISA tests are also more focused on applying knowledge and are more heavily contextualised (ibid).

possible explanation for this is that the standards measured by PISA have changed between 2000 and 2003. Another is that the Key Stage 3 and Key Stage 4 test scores are not consistent over the period. Our preferred explanation is that improvements in the key stage results do not accurately mirror improvements in underlying pupil attainment, and that some of the improvement is due to more narrowly focused teaching.

Does the current system of assessment test the necessary range of skills and abilities? While the core subjects of English, mathematics and science are extremely important, there is growing evidence that young people need to develop a much wider range of skills such as the ability to work with others, to listen and present effectively, to reflect critically, to stay 'on task' for extended periods of time and to act responsibly. These have sometimes been described as 'soft skills' (non-cognitive skills), but recent work by ippr indicates that these are essential skills for life, that they have been growing in importance, and that there is now a significant social class gap in their attainment (Margo *et al* 2006). Some studies show that non-cognitive skills (personal and social skills and personality attributes) are as important as cognitive abilities (such as reading and mathematics ability at age 10) in determining earnings later in life, and analysis of the 1958 and 1970 cohorts indicates that non-cognitive skills became significantly more important over this period (Blanden *et al* 2006).

While teachers often try to develop these skills in their pupils, it is not at all clear that they are effectively specified in the curriculum and assessed by the current system of national tests. The danger is that the current assessment system thus fails to consider some crucial aspects of a young person's development. We should try to create space for this in a reformed system, but we want to do so in a way that maintains school accountability.

Assessment for accountability

Schools are now held much more strongly accountable for the outcomes achieved by their pupils, and their attainment at the end of the key stages in particular. One of the mechanisms for this is parental choice of school, and we discuss this further in another paper (Brooks and Tough 2006). In addition, the results of national tests are a critical input into Ofsted inspections, and a bad inspection may result in a school being issued a notice to improve, or risk being placed in special measures. Entering special measures means that a school loses its autonomy and represents a severe criticism of the leadership of the school. Failure to emerge from special measures rapidly enough can result in a school being closed entirely. School leaders thus face very clear incentives to ensure that their results do not prejudice their inspection results.

It is quite right that there should be a robust inspection mechanism to provide schools with powerful incentives to improve, and especially to ensure that no school falls below a minimum acceptable standard. However, if test results are to play an important role in such a powerful incentive mechanism, it is all the more important that they are robust, valid, and do not negatively impact on other desirable aspects of the learning environment. This particular issue – that preparation for tests might be crowding out other desirable activities in schools – is deal with in the next section of this paper.

Test results are more reliable at the school level than for individuals, because in a large group of pupils individual misclassifications tend to cancel each other out. However, the problem of validity is equally acute at the school level as it is for individual pupils. Schools are held accountable for their test results. The evidence set out in the previous section of this paper suggests that test results are overstating real improvements in young people's abilities. The danger is thus that we are holding schools accountable for the wrong thing. Another danger is that because non-cognitive skills are not recognised by the assessment system, schools have little incentive to focus on their development.

It is important not to overstate these arguments. Ofsted inspections do take into consideration a wide range of factors in addition to test results. Even if there is a serious question about the validity of the tests, a school that is achieving poor test results, given its intake, is unlikely to be successfully developing the skills those tests are meant to be measuring. However, it is certainly the case that schools do have strong incentives to focus on the results of the tests we currently have. If there are problems with the validity of those assessments, there will be a problem with accountability. What is needed is not less accountability, but more valid and reliable tests.

Assessment for learning

How widespread is good practice in assessment for learning in England? Ofsted finds that the use of assessment for learning is only good or better in a minority of secondary schools and unsatisfactory in a quarter (Ofsted 2005a) and that it is the least successful element of teaching in primary schools (Ofsted 2005b). It concludes that schools often do not seem to have the capacity, training or urgency to process information properly for use within the classroom. Despite efforts by DfES to spread good practice, assessment for learning needs to be given a higher priority in both initial teacher training and continued professional development. Responding to the 2006 General Teaching Council's Annual Survey of Teachers[8], 42.6 per cent of teachers reported that they 'will need' professional development in Assessment for Learning over the next 12 months (Hutching *et al* 2006).

One factor that is often cited as a barrier to effective teaching is an over-prescriptive and over-crowded curriculum. Nearly one third of teachers feel there is 'little' or 'no' flexibility in the curriculum (Hutching *et al* 2006). In fact, the national curriculum is much less restrictive than is often claimed, and schools also have the right to apply to the Secretary of State for exemptions from curriculum requirements.

Does the current system of national tests act as a barrier to the adoption of assessment for learning? A key argument of opponents of the current assessment system is that too much teaching time is taken up with non-productive preparation for tests, and that this is squeezing out other more valuable activities. A first question to ask is whether it is possible that some forms of preparation for tests might represent a bad use of classroom time, or on the other hand whether everything that improves test results is useful.

In fact, there does seem to be a range of activities that are intuitively of low educational value that nonetheless might improve test results:

- **Narrow learning.** Because all pupils take the same test, and because each test covers a relatively large subject area, it is possible to predict with reasonable accuracy what will be tested – the most prominent elements of the curriculum. This enables teachers to concentrate on those aspects of the curriculum that are most likely to come up in the tests.
- **Shallow learning.** Because all pupils take the same kind of test, it is possible to predict with reasonable accuracy how each component of the curriculum will be tested. This enables teachers to focus on this approach.
- **Question spotting.** This is essentially follows from the previous two problems.
- **Risk-averse teaching** with low levels of innovation.

It is very difficult to be sure of the extent and impact of practices such as shallow and narrow teaching and learning, and even more difficult to prove a causal link between their prevalence and the nature of the assessment system. However, there is evidence that teachers' own assessments become less formative and more summative in response to high-stakes testing. The Primary Assessment, Curriculum and Experience (PACE) project, a longitudinal study that followed a cohort of primary school pupils for eight years starting before the introduction of national tests for seven-year-olds, found that after the introduction of the tests teachers' own classroom assessment became more summative (Pollard *et al* 2000).

It is also important to emphasise that we do not have to choose between doing assessment for learning and assessment for summative purposes. It is important to remember that the key objective of assessment for learning is improved child development. There should thus in theory be no tension between formative assessment and evaluative assessment, because the first should be a means to improvement in the second. To some extent what is needed is a jump from one equilibrium, in which teachers are too busy training their pupils to pass national tests to focus on assessment for learning, to a better one in which teachers make more use of assessment for learning and as a result their pupils perform better in national tests.

8. A random sample of 10,000 teachers was drawn from a sample pool of 430,722 eligible teachers registered with the General Teaching Council, that is, those who were in service in state schools in England in September 2005. In total 3665 completed questionnaires were received, a response rate of 37 per cent. The achieved sample was compared with the population in terms of key variables, and while there were minor differences between the sample and the population, these were small enough not to affect representativeness, so that generalisation from the sample to the population could be made with confidence.

Recommendations

We believe that it is possible to change the assessment system so that it facilitates better teaching and learning, provides a better measure of child attainment, and maintains a high level of school and national accountability. In summary, our proposals are:

- Every child should be assessed throughout each key stage by their teachers.
- Every child should sit a small number of national tests at the end of each key stage, but not in every area of every subject. The results of these monitoring tests should be used to measure overall school performance, but not the individual pupil's attainment.
- The school's performance should be used to moderate the teacher assessments, producing detailed, nationally-comparable data for each pupil.

Every child should be assessed throughout each key stage by their teachers

Short tests lead to unreliable results, and written tests can only assess certain kinds of ability. We should therefore use a wider range of assessment methods, for a broader range of skills, over a longer period. All this suggests that evidence of the level of a pupil's ability should be gathered over the length of their study rather than in an hour-long examination, which further implies a bigger role for teacher assessment.

It should be possible in large part to use formative assessment for summative purposes. The TGAT Report (Task Group on Assessment and Testing) concluded in 1987 that formative assessment could 'meet all the needs of national assessment at ages before 16' (V.26). TGAT recommended that formative assessment should be the basis of national assessment at seven, 11 and 14 and that assessment should only be designed for summative purposes at the end of compulsory schooling when information is required for certification. For the seven, 11 and 14 age groups, key elements of their scheme included:

- A significant role for teacher assessment. This would require teachers to rate pupils according to criteria set out for each level in the national curriculum.
- Emphasis on assessing a full range of skills. A range of standardised, cross-curricular assessment tasks would ensure that those skills not easily measurable through written tests were nonetheless assessed.

Such a system would represent a major challenge to the teaching profession, and would be demanding to implement. In 1993, teachers boycotted standardised assessment tasks, complaining that the practical classroom observations and tasks and complex 'criterion reference' marking were too burdensome. There were, for example, 17 attainment targets for science and 15 for maths on which teachers had to pass judgments for each child (Green 2006). If teacher assessment is to be a success then it will need to be designed and implemented in a way that has the support of teachers and is not excessively burdensome to operate, and it will require significant investment in professional training and development.

What can be offered to teachers in return for the challenges of a new assessment system is the opportunity for better, more appropriate and more effective exercise of their professional skills. The evaluation of the trial exploring a move to reporting-only teacher assessment at Key Stage 1 (which was rolled out nationally in 2004) reported that many teachers saw opportunities for reducing their workload as well as for developing their professionalism (Shorrocks-Taylor *et al* 2004). The report also found that parents generally preferred their child not to be tested but noted that parents still wanted information regarding schools' test performance (ibid). Research shows that 70 per cent of head teachers (sample 375) and 76 per cent of Year 2 teachers (sample 306) felt that the new assessment arrangements at Key Stage 1 had a positive effect on teaching and learning in the classroom (Reed and Lewis 2005).

Another major worry about teacher assessment is that it will be unreliable because teachers will not accurately grade their pupils. Part of this concern may be that there would be unintended bias in teacher assessment. The experience of shifting to teacher assessment at Key Stage 1 does not support this hypothesis – national results went down after its introduction in 2004. Part of the concern may also be that teachers will adjust their expectations and therefore their grades in the light of their school's or class's circumstances. All judgments are relative and therefore teacher judgments on individual pupils may be affected by the range of abilities in the group (Laming 2004). If this happened then grades would be inflated in schools with low average attainment, and depressed in schools with high average attainment. A system of monitoring testing and moderation would ensure this would not happen.

Every child should sit a small number of national tests at the end of each key stage, but not in every area of every subject. The results of these monitoring tests should be used to measure overall school performance, but not the individual pupil's attainment

National testing performs two vital functions. First, it provides an independent and external measure of school performance, and second it provides a measure of overall national progress. However, it may be possible to fulfil both of these functions and at the same time reduce the burden of such tests on individual pupils and on the schools system as a whole. The key is to stop using the national tests to measure individual pupil attainment.

For national tests to measure individual pupil performance on a consistent basis, each pupil must sit the same tests and must be tested in every subject. However, if the tests are being used only to measure school and overall national performance, it may be possible for each pupil to sit tests in just some of their subjects, and for different pupils to sit different tests in the same subject. The tests could thus in aggregate cover much more of each subject, and it would become much more difficult for teachers (and pupils) to predict their content. This would make it much harder to 'teach to the test', and even if there would still be some scope to narrow the focus of learning due to the inherent limitations of timed tests, it would become necessary to teach to the curriculum more broadly.

These tests could be used in the same way as the current key stage tests to hold schools to account for their pupils' performance. Schools would thus still have a strong incentive to achieve good results, even though these would not directly determine individual pupils' scores. Shifting towards a system of monitoring tests would be independent of any decision about the publication of school performance information. Once school performance data is collected it can be collated and presented in table format, so even if DfES stopped publishing performance tables it is likely that it would continue to be produced. Monitoring test data could be in the form of absolute performance, value added, or contextualised value-added measures. At the school level very much the same data would continue to be available as it is now. Accountability would if anything be improved because the results would become a more valid reflection of true pupil abilities across a wider curriculum. At the national level the test results would give a more valid and reliable picture of pupil attainment, because they would cover the curriculum in a much broader and deeper way than is possible where every student takes the same exam.

In the current system each pupil is tested on each subject area, namely English, mathematics and science. In the proposed system all that would be necessary is a monitoring test that gives a statistically reliable guide to the overall performance of the school. The required number of tests for each pupil would thus vary with the size of the school. For a large school this might significantly reduce the amount of national tests that each pupil would have to sit compared to the status quo. For a very small school it might not be possible to reduce the number very much. However, it is important to remember that even the current extensive system of tests does not provide a reliable guide to the quality of small schools, whose results can fluctuate significantly from one year to the next simply due to the small number of students being tested. Another concern might be that although schools have strong incentives to achieve good results, pupils do not, and they might therefore not put any effort into the tests. This may or may not be a significant issue: arguments can be made in theory either way, and more research and evaluation will be required in this area.

The school's performance should be used to moderate the teacher assessments, producing detailed, nationally comparable data for each pupil

The final piece of the puzzle is to use the monitoring tests to moderate teacher assessments. A system of unmoderated teacher assessment would be unlikely to command public confidence, as it would be possible for different teachers to be allocating different grades to similar pupils.

Part of the response to this problem should be to try to improve the quality of teacher assessment for both formative and evaluative purposes, both in initial teacher training and in continued professional development. This is likely to be an important part of any major shift in this direction. The evaluation of the Key Stage 1 trial found that 'accuracy and teacher confidence in making Teacher Assessments is strongly affected by the quality of training and moderation' (Shorrocks-Taylor *et al* 2004: 4). Assessment for learning should certainly be given significantly more profile in initial teacher training. At present, formative assessment is not mentioned explicitly in the professional standards for teachers. The standards are currently under review and the draft revised standards for Qualified Teacher Status (the standards that need to be reached to become a qualified teacher) do include specific reference to being 'informed of … the importance of formative assessment' as well as to 'know how to use local and national statistical

information to evaluate the effectiveness of their teaching, to monitor the progress of those they teach and to raise levels of attainment' (TDA 2006, Q11: 10).

It would also be possible to develop of a cadre of specialist teacher assessors, to encourage the exchange of teachers involved in assessment between different schools, or to develop professional external moderators to assist schools. One option to raise the status of teachers and their ability to undertake accurate assessment would be to have at least one qualified assessor in each school. This idea was first put forward by the Association of School and College Leaders (ASCL) (and then the Secondary Heads Association), which described a vision of a chartered examiner in every large department in secondary school.

There are many advantages to such a model. It could help to restore trust in teacher assessment, and it would also provide a route similar to the 'excellent teacher' or 'advanced skills teacher': a route of progress for experienced teachers who do not want to take the leadership route. There would be increased costs involved as these assessors would command a higher salary as well as more staff time dedicated to preparing for and moderating assessment. However, as with the costs associated with the increased training, these would be balanced with the reduced financial burden of the external examinations bill. Recent research commissioned by the QCA indicates the total cost of the examinations and testing system in England to have been £610 million in 2003-04 (QCA/PwC 2004). A more accurate reflection of the costs of National Curriculum Testing would be £346 million as this removes the direct costs of the three awarding bodies for post-14 qualifications (Edexcel, AQA and OCR)[9].

However, the best guarantee of comparable results and thus high levels of public confidence would be, in addition to any other measures, to use the monitoring test results to adjust the teacher-assessed pupil scores. We have not worked through the details of such a moderation scheme, and there would undoubtedly be complexities. However, the schematic picture is as follows: the monitoring tests independently indicate the distribution of attainment in the school in each of the core subjects. This information can then be used to scale the teacher assessments for individual pupils so that they fit into the known school-level distribution.

Conclusions

A system of assessment such as the one outlined here would require extensive research and development, piloting and phased introduction alongside a major programme of teacher training and the creation of new systems of teacher assessment and national testing. We do not claim to have a model, but have presented here the broad features of such a system. It would represent a revolution for English education, but could potentially meet each of our objectives better than the existing system of assessment:

- It should be more reliable, because teacher assessments could be based on much more information than can be captured in a single round of tests.
- It should be more valid at the national level as there would be data on a wider range of subject areas.
- It should have greater validity at the pupil level, because teacher assessments could more successfully measure the different aspects of each pupil's progress.
- It should thus provide a better measure of pupil attainment, beyond pen and paper tests.
- The monitoring tests should maintain accountability at the school level, and should provide a better measure of national progress on standards.
- It should facilitate assessment for learning, both because teacher assessments of individual pupils could be built up from formative assessment results, and because it would make it much more difficult to 'teach to the tests' and should thus promote teaching the whole curriculum in the most effective way.
- Teachers would have more time to focus on the development of other important skills such as non-cognitive skills.

9. Although this figure still includes the costs of administering GCSEs, A Levels, and so on, for exam centres.

References

Note: web references correct at December 2006

Black P and Wiliam D (1998) *Inside the Black Box: Raising Standards Through Classroom Assessment* London: King's College London

Black P and Wiliam D (1998a) 'Assessment and classroom learning' *Assessment in Education* 5 (1) 7-74

Black P and Wiliam D (2006) 'The Reliability of Assessments' in Gardner J (ed) (2006) *Assessment and Learning* London: Sage

Blanden J, Gregg P and Macmillan L (2006) *Accounting for Intergenerational Income Persistence: Non-Cognitive Skills, Ability and Education* London: Centre for the Economics of Education, London School of Economics, reported in Margo J and Dixon M with Pearce N and Reed H (2006) *Freedom's Orphan's: Raising youth in a changing world* London: Institute for Public Policy Research

Bloom B, Hastings T and Madaus G (1971) *Handbook of Formative and Summative Evaluation of Student Learning* New York: McGraw-Hill

Brooks R and Tough S (2006) *Pupil Attainment: Time for a three Rs guarantee* London: Institute for Public Policy Research. Available at: www.ippr.org/publicationsandreports

Daugherty R (1995) *National Curriculum Assessment: A Review of Policy 1987-1994* London: Falmer Press

Department for Education and Skills (DfES) (2006a) 'National Curriculum Assessments at Key Stage 1 in England, 2006 (Provisional)' *Statistical First Release* 30/2006 24 August

DfES (2006b) 'National Curriculum Assessments at Key Stage 2 in England, 2006 (Provisional)' *Statistical First Release* 31/2006 24 August

DfES (2006c) 'National Curriculum Assessments at Key Stage 3 in England, 2006 (Provisional)' *Statistical First Release* 34/2006 13 September

Gardner J and Cowan P (2000) *Testing the Test: A Study of the Reliability and Validity of the Northern Ireland Transfer Procedure Test in Enabling the Selection of Pupils for Grammar School Places* Belfast: Queen's University of Belfast

Gardner J and Cowan P (2005) 'The fallibility of high stakes '11-plus' testing in Northern Ireland' *Assessment in Education* 12 (2): 145-165

Green S, Johnson M, O'Donovan N and Sutton P (2003) 'Changes in Key Stage Two Writing From 1995 to 2002' A paper presented at the United Kingdom Reading Association Conference at University of Cambridge, 11-13 July

Green S (2006) in Reed J and Tough S (eds) (2006) *Curriculum, Assessment and Pedagogy: Beyond the "standards agenda"* London: Institute for Public Policy Research

Hutchings M, Smart S, James K and Williams K (2006) *General Teaching Council for England Survey of Teachers 2006* London: Institute for Policy Studies in Education, London Metropolitan University. Available at: www.gtce.org.uk/research/tsurvey/tsurvey06/

Laming D (2004) *Human Judgement: The Eye of the Beholder* London: Thomson

Margo J and Dixon M with N Pearce and H Reed (2006) *Freedom's Orphan's: Raising youth in a changing world* London: Institute for Public Policy Research

Massey A, Green S, Dexter T and Hamnett L (2003) *Comparability of national tests over time: Key stage test standards between 1996 and 2001: Final report to the QCA of the Comparability Over Time Project* Cambridge: University of Cambridge Local Examinations Syndicate. Available at: www.qca.org.uk/6301.html

Micklewright J and Schnepf S V (2006) 'Response Bias in England In PISA 2000 and 2003' *DfES Research Report* RR771. Available at: www.dfes.gov.uk/research/

Miliband D (2003) 'School Improvement And Performance Management' A speech by David Miliband to Perfomance Management Conference, Bournemouth, 27 January. Available at: www.dfes.gov.uk/speeches/search_detail.cfm?ID=59

Ofsted (2005a) *The Secondary National Strategy: An evaluation of the fifth year* Reference No. HMI 2612. Available at: www.ofsted.gov.uk/assets/4118.pdf

Ofsted (2005b) *Primary National Strategy: An evaluation of its impact on primary schools 2004/2005* Reference No. HMI 2396. Available at: www.ofsted.gov.uk/assets/4117.pdf

Ofsted (2006) *Evaluating mathematics provision for 14-19 year olds* Reference No. HMI 2611. Available at: www.ofsted.gov.uk/assets/4207.pdf

Pollard A and Triggs P with Broadfoot P, McNess E and Osborne M (2000) *What pupils say: changing policy and practice in primary education – findings from the PACE project* London and New York: Continuum

QCA/PwC (2004) *Financial Modelling of the English Exams System 2003-04* London: QCA. Available at: www.qca.org.uk/12130.html

Reed M and Lewis K (2005) *Key Stage 1 Evaluation of New Assessment Arrangements* London: NAA. Available at: www.qca.org.uk/downloads/pdf_05_18931.pdf

Richards C (2005) *Standards in English Primary Schools: are they rising?: A contribution to the debate from the Association of Teachers and Lecturers* London: Association of Teachers and Lecturers. Available at: www.askatl.org.uk/atl_en/resources/publications/research/Primarystandards.asp

Ruddock G, Sturman L, Schagen I, Styles B, Gnaldi M and Vappula H (2004) *Where England stands in the trends in international mathematics and science study (TIMSS) 2003: Summary of national report for England* Slough: NFER. Available at: www.nfer.ac.uk/publications/other-publications/downloadable-reports/where-england-stands-in-the-trends-in-international-mathematics-and-science-study-timss-2003.-national-report-for-england.cfm

Ruddock G, Clausen-May T, Purple C and Ager R (2006) 'Validation Study of the PISA 2002, PISA 2003 and TIMSS-2003 International Studies of Pupil Attainment' *DfES Research Report* RR772. Available at: www.dfes.gov.uk/research/

Shorrocks-Taylor D, Swinnerton B, Ensaff H, Hargreaves M, Homer M, Pell G, Pool P and Threlfall J (2004) *Evaluation of the trial assessment arrangements for key stage 1: Report to QCA* London: QCA. Available at: www.qca.org.uk/8994.html

Statistics Commission (2005) 'Measuring Standards in English Primary Schools: Report by the Statistics Commission on an article by Peter Tymms' *Research Report* 23 London: Statistics Commission. Available at: www.statscom.org.uk/C_402.aspx

Training and Development Agency (TDA) (2006) *Draft revised standards for teachers*. Available at: http://www.tda.gov.uk/teachers/professionalstandards.aspx

TeacherNet (2006) National curriculum assessment arrangements. Available at: www.teachernet.gov.uk/management/atoz/n/nationalcurriculumassessmentarrangements/

Tymms P (2004) 'Are standards rising in English primary schools?' *British Educational Research Journal* 30:4 477-494

Wiliam D (2000a) 'The meanings and consequences of educational assessments' *Critical Quarterly* 42(1)

Wiliam D (2000b) 'Reliability, validity and all that jazz' *Education* 29 (3)