British Journal of Educational Studies

ISSN: 0007-1005 (Print) 1467-8527 (Online) Journal homepage: http://www.tandfonline.com/loi/rbje20

The Return to Final Paper Examining in English National Curriculum Assessment and School Examinations: Issues of Validity, Accountability and Politics

Harry Torrance

To cite this article: Harry Torrance (2018) The Return to Final Paper Examining in English National Curriculum Assessment and School Examinations: Issues of Validity, Accountability and Politics, British Journal of Educational Studies, 66:1, 3-27, DOI: 10.1080/00071005.2017.1322683

To link to this article: https://doi.org/10.1080/00071005.2017.1322683

Published online: 18 May 2017.

British Journal of Educational Studies
Vol. 66, No. 1, March 2018, pp. 3–27

THE RETURN TO FINAL PAPER EXAMINING IN ENGLISH NATIONAL CURRICULUM ASSESSMENT AND SCHOOL EXAMINATIONS: ISSUES OF VALIDITY, ACCOUNTABILITY AND POLITICS

by HARRY TORRANCE, *Manchester Metropolitan University, Manchester*

ABSTRACT: There are sound educational and examining reasons for the use of coursework assessment and practical assessment of student work by teachers in schools for purposes of reporting examination grades. Coursework and practical work test a range of different curriculum goals to final papers and increase the validity and reliability of the result. However, the use of coursework and practical work in tests and examinations has been a matter of constant political as well as educational debate in England over the last 30 years. The paper reviews these debates and developments and argues that as accountability pressures increase, the evidence base for published results is becoming narrower and less valid as the system moves back to wholly end-of-course testing.

Keywords: final paper examinations, English National Curriculum tests and examinations, validity and reliability of tests and examinations, coursework assessment, the politics of assessment

1. INTRODUCTION: STABILITY AND CHANGE IN ASSESSMENT POLICY AND PRACTICE

The involvement of teachers in the formal assessment of their students for examination purposes has a long history in England. Teacher assessment or school-based examining as it is sometimes known was envisaged in the Norwood Report (1943, preceding the 1944 Education Act) and recommended by the Beloe Report (1960) which led to the setting up of the Certificate of Secondary Education (CSE) in 1963. Teacher assessment became a significant element of CSE, started to be used in the General Certificate of Education (GCE) O-level, and was a key feature of early pilots and the final specifications of the General Certificate of Secondary Education (GCSE) which superseded CSE and O-level in 1986 (DES, 1982; Waddell Report, 1978). As the Waddell Report put it:

> assessment over a period of time by the teacher who knows the pupil… was found useful in searching out skills and understanding which may be more readily tested in this way than in a formal written examination… The teacher can observe and

ISSN 0007-1005 (print)/ISSN 1467-8527 (online)

https://doi.org/10.1080/00071005.2017.1322683
http://www.tandfonline.com

> assess the way in which the pupil sets about the process of solving practical problems in science or… how he develops a response to criticism in activities such as writing. (paragraph 27)

While there have always been debates about the extent to which teacher assessment could be regarded as valid and reliable, and how best to achieve this (Cohen and Deale, 1977; Hoste and Bloomfield, 1975), it was not until considerably after the introduction of GCSE in 1986 (with first examinations in 1988) that the practice became so politically controversial.

Only 2 years after the introduction of GCSE, the National Curriculum and National Testing system was introduced in 1988. Following initial policy statements and legislation (DES, 1988), the system that emerged in the early 1990s was nowhere near as ambitious as originally envisaged, which would have involved setting up national tests in every curriculum subject (up to 10). Nevertheless, the new system still involved testing all children at ages 7, 11 and 14 in English, Maths and Science, coupled with the use of national end-of-school examinations, the GCSE, taken in a range of individual subjects, to form the 'national test' programme at age 16. Moreover, these tests involved a combination of internal school-based teacher assessment and externally set and marked tests so the operational scale and scope of the system was immense. Teacher assessment was used along with externally set and marked tests in order to maximise the validity, reliability and utility of the system (DES, 1988; see Daugherty, 1995; Torrance, 2003; for a longer account of developments).

The original purposes of the new National Curriculum Assessment system were set out as diagnostic, formative, summative and evaluative (DES, 1988) – thus attempting to monitor the progress of individual pupils as well as the standards attained by the system as a whole. Subsequently, however, the summative and evaluative purposes of national testing have become more and more prominent as moves towards greater accountability of schools and teachers gathered pace (Stobart, 2001, 2009). Now National Curriculum Assessment is seen primarily as ensuring the accountability of the school system with enormous focus on the publication of national aggregate results and results for cohorts at the level of the individual school. Such results are also the focus of inspection visits and reports. At one and the same time, the scale and complexity of the system has led to successive scaling back of the range of assessments employed and the age cohorts involved. External testing at age 7 was dropped in 2004 after extensive criticisms of testing very young children, though externally designed tests are provided to teachers to administer to 7-year-olds 'to help them to reach an overall judgement of the standards children have reached in English reading and mathematics' (STA, 2016a, p. 3). Results are reported to parents and available to Office for Standards in Education (OfSTED) inspections. Debate continues over whether or not the government will reintroduce external testing for 7-year-olds (Guardian, 2015; Morgan, 2015a; NUT, 2016). Testing at age 14 was abandoned in 2008 after a fiasco of lost papers and missing results during that summer's round of testing (Torrance, 2009). Science testing at age 11 was ended in

2010 in a further effort to reduce the logistical burden on individual schools and the system as a whole (Isaacs, 2010; Porter, 2008; Turner, 2009).

National Curriculum Assessment now involves testing English and Maths at age 11, along with the use of GCSE at age 16. Paradoxically, therefore, as the significance of National Curriculum Assessment has grown for purposes of accountability, the evidential base for this purpose has been narrowed – a key issue for this paper. Effectively, the accountability of English primary schools rests on externally set and marked tests of English and Maths at age 11. Until 2015, secondary schools were measured by the number of students gaining good (passing) grades in at least five GCSEs at age 16. From 2016 results onwards, a value-added measure known as 'Progress 8' is being used – essentially a measure of whether or not students have met or exceeded expectations of GCSE grades based on performance at age 11 (DfE, 2017). The calculations are very complex, however, and depend upon the validity of the assessments at age 11 (Key Stage 2). It remains to be seen whether this replaces 'five good A-Cs' in public understanding and discourse, or its equivalent of grades 9–5 under the new grading system to be introduced in 2017. Speaking in 2015, the then Secretary of State, Nicky Morgan focussed on grade 5 as 'the new "good pass" that will be used to hold the government and schools to account' (Morgan, 2015b).

2. Validity in test construction and use

Validity of assessment processes and outcomes can be analysed with respect to the accuracy and fairness of the measurement of individual achievement and with respect to the uses to which the measurement is put. The correct – i.e. valid and reliable – measure could still be used inappropriately. So it is with using measures of student performance to hold teachers and schools to account via national assessment systems. Extensive discussions of validity can be found in the assessment literature detailing aspects of content, construct, concurrent, predictive and consequential validity (e.g. Messick, 1989). However, for the purposes of the discussion here the issue basically revolves around content and construct validity on the one hand – is the test measuring what it purports to measure; and consequential validity on the other – how is the test used and are the measures involved appropriate? Consequential validity is a contested idea in the assessment literature (Newton and Shaw, 2016). Some argue that while it is important to know about and evaluate the uses to which test is put, these are not strictly matters for the validity of the measurement per se. The validity of a test should be restricted to the properties of the test itself (Popham, 1997). However, given that contemporary national testing regimes are now largely developed for purposes of accountability, consequential validity must surely come into the discussion. Newton and Shaw (2016) indicate that validity is now generally understood as an integrated and unified concept and this shift in test use from measurement tool to accountability policy lever has also been noted internationally with similar implications for discussions of validity (OECD, 2013; Weiner, 2013). Both interpretations of validity are addressed in the paper.

With respect to the development of assessment instruments, particularly with respect to individual measurement, validity requires appropriate sampling of curriculum goals and content and of the varying circumstances under which knowledge is demonstrated and achievement is produced – coursework, oral work, practical work, fieldwork and so forth – in addition to test-based recall of knowledge. This is likely to involve the use of school-based assessment conducted by teachers (teacher assessment) in order to increase the sample of work that can be assessed (Johnson, 2013; Stanley *et al*., 2009). Indeed, these are the arguments that led to teacher assessment being introduced in CSE in the 1960s and to GCSE in the 1980s. Having said this, however, as noted earlier, it is also the case that there has been extensive debate and investigation around the extent to which teachers can reliably assess their own students. The validity of an assessment also depends on the assessment being reliable i.e. that it is accurate, at least within acceptable tolerances. Reliability is usually defined in terms of replicability – would the same mark be awarded for the same work on a different occasion, and/or by a different marker. Studies indicate that teachers can rank order within a cohort (i.e.in their own school/classroom) fairly reliably, but may be out of synch with standards in other schools, and this is often the focus of attention for moderation systems (Black *et al*., 2010; Harlen, 2005; Stanley *et al*., 2009). Reliability is often privileged over validity in assessment but more nuanced discussions recognise that they are inextricably intertwined (Baird and Black, 2013; Newton and Shaw, 2016). Threats to reliability not only can derive from conscious and unconscious bias (e.g. with respect to race, social class and gender, Gipps and Murphy, 1994; Gipps and Stobart, 2009) but also from teachers not being sufficiently aware of practice and levels of achievement elsewhere. Research suggests that despite the arguments and decades of policy and practice, there is actually rather limited evidence concerning the validity and reliability of teacher assessment (Johnson, 2013). Research also suggests, however, that teachers *can* assess their students reliably, particularly when working with strong moderation processes across schools, but that this involves considerable investment in professional development and moderation procedures (Black *et al*., 2010, 2011; Stanley *et al*., 2009). I will return to this issue later in the paper.

With respect to system accountability, validity similarly depends on appropriate sampling of the full range of outcomes that policy is seeking to pursue. Evaluating the performance of individual schools and indeed a whole school system on the basis of a small number of tests is clearly inadequate. Validity also depends on limiting the 'washback' effect on teaching and learning, since practising and coaching for tests can render the results invalid. Not only can practising for tests narrow the curriculum that is taught, but also the construct that is being measured becomes test preparation rather than English and/or Maths. Yet this is exactly the impact that National Curriculum Assessment has had on teaching in England (Cambridge Primary Review, 2010; Torrance, 2011).

Even the government's own school Inspectorate and the House of Commons Education Committee have noted such implications with alarm:

> In many [primary] schools the focus of the teaching of English is on those parts of the curriculum on which there are likely to questions in national tests. . .History and, more so, geography continued to be marginalized . . . In [secondary] schools. . . the experience of English had become narrower. . . as teachers focused on tests and examinations. . . There was a similar tension in mathematics. . .(OfSTED, 2006, pp. 52–56)

> In an effort to drive up national standards, too much emphasis has been placed on a single set of tests and this has been to the detriment of some aspects of the curriculum and some students. (House of Commons Select Committee reported on BBC 13 May 2008, http://news.bbc.co.uk/1/hi/education/7396623.stm)

3. Validity and the wider goals of education

In addition to what we might call the more technical issues of validity, the scope, scale and ambition of assessment have changed enormously over the last 30 years or so, thus also changing the nature of the debate over validity and its relationship to accountability. Assessment was originally developed to identify individual achievement and select a small number of elite students for further study. Now, however, both the intellectual field of assessment theory and practice, and the education policy context, assume that all, or at least the overwhelming majority, of the school population can and should be educated to the highest level possible (Broadfoot, 1996; Hargreaves, 1989; Horton, 1988). The focus of policy is now on education for all and the development of a fit-for-purpose assessment system *as a system*, i.e. as part of an integrated approach to national human resource development. The imperative now is to treat education as an economic investment, both on the part of the individual student and on the part of government (Brown and Lauder, 1992; Lauder *et al*., 2006; Wolf, 1995). Elsewhere I have characterised these developments in assessment as a move from a 'technology of exclusion' to a 'technology of inclusion' (Torrance, 2016). Assessment is now expected to accurately identify and report the individual educational achievement of the vast majority of the student population; in turn, such measures are also expected to accurately evaluate the effectiveness of individual schools and national school systems, and also sometimes individual teachers (OECD, 2013).

Moreover, the range of what is taught and assessed has expanded. Employers, university admissions tutors and so forth want more information about what school leavers can do, and governments want more information about what the school system is producing. Demands have also grown for the school system to produce different things – a wider range of more relevant skills and understandings for the so-called 'knowledge economy' (Hargreaves, 1989, 1994; Scardamalia *et al*., 2012). This in turn has required the development of a wider range of assessment methods to identify and report a wider range of learning outcomes – practical work, coursework and extended project work, for

example, to test practical competences and the application, rather than simply the memorisation and regurgitation of knowledge. Thus, a concern for what we might term 'content standards', and the production of more useful information about what school students know, understand and can do, has merged with debates about how best to measure and report such content standards and indeed enforce them. Additionally, assessment – formative assessment – is also now expected to support and underpin the process of learning, not just to measure the outcomes of learning, and so, again, the expectations for the field are vastly more ambitious than was once the case.

This expansion of the scope and purpose of assessment means that there are good reasons, rooted in traditional assessment concerns for validity and reliability, for the development of more coursework, project work and so forth in national test systems and school examinations, i.e. for an increased role for teachers in setting and marking examination work in their own schools and classrooms. Validity demands that the pursuit of broader curriculum goals such as analysing data, applying knowledge and developing practical skills is underpinned by broader methods of assessment. These new skills and abilities cannot be tested by written final examinations alone. Equally, reliability demands that these and other skills and abilities should not simply be measured by a one-off test, but assessed on several occasions over a longer period of time: the argument being that the larger the sample of assessed work, undertaken under a variety of conditions, the more reliable the result is likely to be. Moreover, as a recent OECD report has noted:

> The central agent in securing links between the evaluation and assessment framework and the classroom is the teacher. This highlights the importance… [of drawing] on the professionalism of teachers in ensuring…authentic improvement in classroom practices and student learning… (2013, p. 12)

However, evidence from England indicates that this is extremely difficult to achieve in practice with key tensions arising from the accountability pressure to use results to evaluate schools and teachers. Validly and reliably measuring a wider range of learning outcomes by using a wider range of assessment methods sit very uneasily with pressure to use results to evaluate the efficiency and effectiveness of the system. One of the key problems of the debate over these matters in England has been around the role of school-based teacher assessment in producing valid assessment data. The debate has been dominated by political arguments over the assumed lack of reliability of such assessment, rather than educational arguments over how best to maximise validity and reliability across curriculum goals, and it is to these matters that I now turn.

4. Developing a new system

The new GCSE preceded the introduction of the National Curriculum by a couple of years. A major part of the reform was the use of far more coursework,

practical work and project work assessment, to be set and marked in the school. The rationale was, as noted above, that the curriculum needed to change and assessment methods needed to change to match the new curriculum. A wider range of skills, capabilities and understandings needed to be developed for the emerging knowledge economy. This in turn required a wider range of assessment methods to be developed to identify and report a wider range of learning outcome. As the then Conservative Secretary of State for Education Sir Keith Joseph put it:

> we need a wide range of instruments... because assessment has many important aims and we cannot expect a single form of assessment to encompass them all equally well. (Joseph, 1986, p. 180)

Similar arguments were used to underpin the use of a mixture of external testing and internal teacher assessment in the emerging National Curriculum Assessment framework (DES, 1988).

At the same time, interest was also developing in making the curriculum more standards-based and assessment more criterion-referenced. Norm-referenced, rank-ordered grades do not communicate what students have actually achieved, only that some students have done better (or worse) than others. Thus, again, as noted above, a concern for 'content standards', and the production of more useful information about what school students know, understand and can do, merged with debates about how best to measure and report such content standards. The corollary of criterion referencing, however, is that if the criteria are achieved, the grade or certificate is awarded – potentially to everyone – and this carries implications for those more used to seeing the purpose of assessment as discriminating between candidates for selection and the maintenance of educational standards over time.

Almost as soon as GCSE and National Curriculum Assessment were introduced, the new methods of assessment came to be criticised by successive Conservative politicians and governments in the 1990s. Teachers complained of overwork (too much school-based assessment), and Conservative politicians complained that standards were falling as more students started to pass GCSE and in due course the new National Curriculum tests (see below, Figures 1 and 2). The result was that Conservative government ministers following Sir Keith Joseph began to insist that the amount of coursework and practical work included in examinations and national assessments should be limited. Their argument was that coursework and practical work were a burden on teachers and were making the examinations easier to pass. Government ministers stated in correspondence with the School Examinations and Assessment Council that they wanted 'terminal written examinations' to be developed for the national tests at 14 (Daugherty, 1995 p. 52). They also insisted that all GCSE examinations should include no

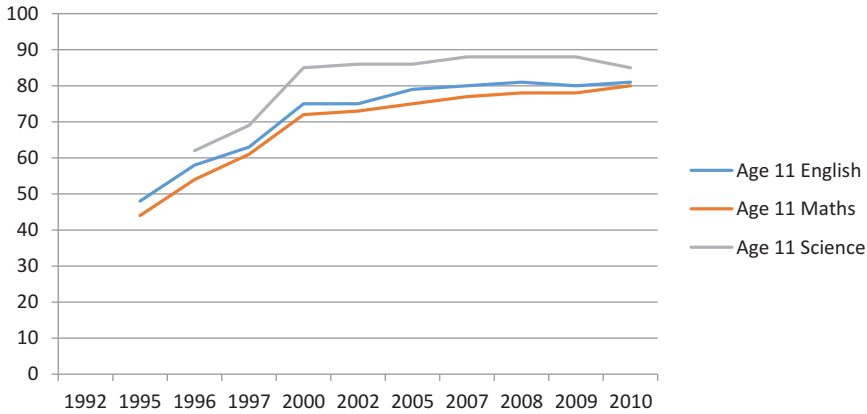

Figure 1. % pupils gaining National Curriculum Assessment level 4 or above at age 11 (KS2), England

Figure 1 finishes with 2010 since this is the last year Science was assessed, and successive different versions of English since 2010 make results difficult to compare. Interestingly, however, only 53% of pupils met the 'new expected standard in reading, writing and mathematics' in 2016, with 70% meeting the standard in maths and 72% in grammar, punctuation and spelling (DfE, 2016a), well below the 80% achieved in 2010.

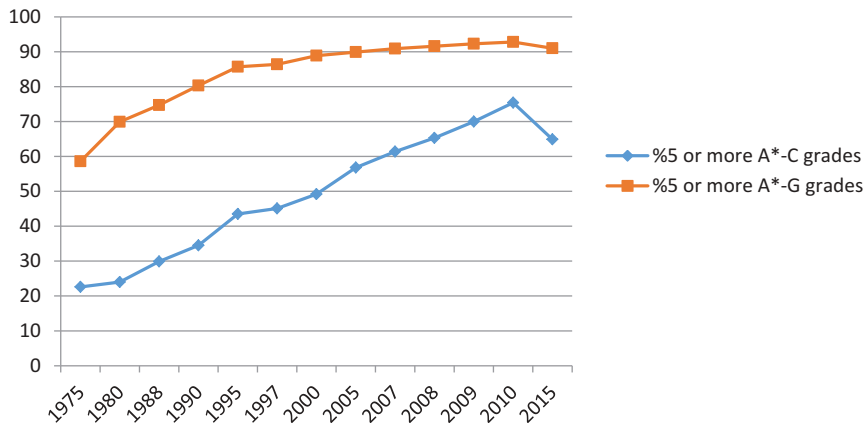

Figure 2. % of pupils gaining O-level/CSE grade 1/GCSE and equivalents 1975–2015, England

more than 20% coursework assessment, and that a proportion of marks should be allocated in all GCSE subjects for spelling, punctuation and grammar (Daugherty, 1995; Dearing, 1994). In a speech to the Centre for Policy Studies in July 1991, the then Prime Minister, John Major, who had succeeded Mrs. Thatcher, stated:

> It is clear that there is now far too much coursework, project work and teacher assessment in GCSE. The remedy surely lies in getting GCSE back to being an externally assessed exam which is predominantly written. I am attracted to the idea that for most subjects a maximum of 20 per cent of the marks should be obtainable from coursework. (quoted in Daugherty, 1995, p. 137)

5. Improving results undermines the credibility of those results

Scroll forward 25 years and the current, UK Conservative government is again reducing coursework and practical work in national testing and school examinations, as will be described below. This time the Conservatives have the convenient excuse that it was the previous Labour government that allowed too many flexible teacher-assessed elements back into school examinations, thereby, they argue (again), lowering educational standards and inflating pass rates. Labour was in power from 1997 to 2010 and allowed the reintroduction and further expansion of coursework and modular assessment of various sorts, as they returned to education policies designed to develop a wider range of skills and competencies for the twenty-first-century knowledge economy. Labour also greatly increased the pressure of accountability on schools by publishing extensive 'league table' data on school performance – i.e. National Curriculum test and examination results, school by school.

Figures 1 and 2 illustrate that in the intervening years from the mid-1990s to the late 2000s pass rates did indeed rise for National Curriculum Assessment and GCSE, though this happened under both Conservative and Labour governments, and in fact the trend of rising secondary school examination pass rates dates back to the 1970s. With respect to National Curriculum tests at age 11 (Figure 1), progress since 1997 was the measure routinely deployed at the national level by the New Labour government (elected in 1997). But there were significant improvements in results prior to 1997. Thus, for example, in the 2 years after National Testing was first introduced at age 11 under a Conservative government (1995–1997), results improved by 15 percentage points in English and 17 percentage points in Maths. In the *10* years after 1997, under Labour, results improved by 16 percentage points in English and 14 percentage points in Maths (1997–2006), with most of this improvement being achieved by 2000 and maintaining a plateau thereafter. One inference we might take from these figures is that the introduction of National Testing constituted a major perturbation in the primary school system, such that results started low, improved rapidly as teachers and students came to understand what was required of them in terms of test preparation and then tailed off as the limits of such artificial improvement were reached. Similar disruption caused by changes in assessment systems and procedures has been previously noted and discussed in the research literature, with identification of what has been called the 'sawtooth effect' whereby:

> performance on high stakes assessments is often adversely affected when that assessment undergoes reform, followed by improving performance over time as students and teachers gain familiarity with the new test. This pattern reflects changes in test-specific performance over time, whilst not necessarily reflecting changes in a cohort's overall mastery of the subject. (Ofqual, 2016a, p. 6)

This effect can also be observed with respect to recent changes to GCSE (see Figure 2 and discussion below).

As regards GCSE (Figure 2), in the mid-1970s, when only the 'top 20%' of students were thought capable of passing the old O-level examination, the percentage of students passing at least five O-levels or their equivalent under the previous dual system was 22.6%.[1] By 1988, the first year of GCSE results, this had risen to 29.9%. By the mid-1990s, this had risen further to 43.5% giving rise to the first wave of Conservative concerns about grade inflation over 20 years ago. Results for 2010, the last year of the 1997–2010 Labour government, and the year when the new Conservative government came to power, indicate that just over 75% of students were passing five or more GCSEs or their equivalent at grades A*–C. That is, by 2010, 75% of the school population were achieving what 30 years previously was thought could only be achieved by the 'top 20%'. Furthermore, taking the full range of grades into account (A*–G), as an indicator of the numbers of students gaining at least some benefit from their secondary education, almost 60% gained at least five A*–G grades in 1975, while nearly 93% achieved five A*–G in 2010.[2]

Pass rates at GCSE since 2010, under the Conservative government, have levelled off and turned down quite significantly as the numbers of vocationally oriented 'GCSE equivalent' examinations have been restricted and political pressure has been exerted on examination boards to address grade inflation. This involves the use of a 'comparable outcomes' approach to establish grade boundaries. Essentially, this involves trying to compare results over time with reference to the previous achievement of the cohort and the mix and number of the students taking each subject (Ofqual, 2015, 2016b). This is being characterised as 'inflation proofing' and bringing 'stability' to the system (Gibb, 2015). Thus, in 2015 only 65% of students secured five or more GCSE A*–C grades. The proposed moves (described below) to final paper-only examinations, now called linear examinations, with the exam at the end of two years, have yet to impact on results but are likely to lead to further reductions, at least in the first instance, as the change impacts on teaching methods and student learning and performance. To divert attention away from these falling pass rates, and establish 2010 as a new baseline for comparisons, the Conservative government is now highlighting the numbers gaining five GCSEs including English and Maths (53.8% in 2015, as against 53.5% in 2010). A completely new measure called the English Baccalaureate (EBacc) has also been introduced comprising five GCSEs including English, Maths, Science, Humanities and Foreign Language. Twenty three per cent of students sitting GCSEs achieved this in 2015 as against

15% in 2010. Future GCSE examinations will be graded from 9 to 1 rather than A*–G, making comparisons over time even more difficult. Grade 9 will be the highest grade, with 1 the lowest, which seems counter-intuitive and may cause some confusion in itself (the old GCE O-level was graded 1–9). Grades 9–1 come into effect from 2017 in English and Maths, and 2018 for other subjects. As noted above, the previous Secretary of State for Education, Nicky Morgan, is on record as stating the new grade 5 will be 'the new "good pass"... comparable to a low B or high C under the old grading system' (Morgan, 2015b p. 1). Even more confusingly, her replacement as Secretary of State, Justine Greening, wrote to the Chair of the House of Commons Education Select Committee in March 2017 that:

> Rather than reporting on the 'good pass', we will instead distinguish between a grade 4 as a 'standard pass' and a grade 5 as a 'strong pass' and report on both. Under the new system, a grade 4 and above will be equivalent to a C and above. (Greening, 2017, p. 1)

It remains to be seen whether or not this becomes the norm in public debate, rather than the new official measure of 'Progress 8' (DfE, 2017).

Nevertheless, currently, we remain in a situation whereby c. 65% of the secondary school population achieve what 40 years ago was thought to be achievable only by the top 20%. Given that these upward trends in results have extended over so many years, some element of a genuine rise in educational standards is likely to be present, driven by better socio-economic conditions of students, higher expectations of educational outcomes by students, parents and teachers alike, and better teaching underpinned by better training and availability of resources. With increased prosperity most of us are living longer; it is therefore at least plausible that most of us are becoming better educated as well. Research over many years demonstrates that measures of educational achievement are positively correlated with socio-economic status (Coleman *et al.*, 1966, Perry and Francis, 2010; Serin, 2005). More recently, however, this trend has been combined with and compounded by two key elements of the changes which have taken place in the system of assessment and accountability:

(1) there is an increasingly more focused concentration on passing examinations, both by teachers ('teaching to the test') and by the majority of students (extrinsic motivation), because of the perceived importance of educational success in institutional accountability, teacher career progression and student life chances;
(2) at the same time increased transparency of modular, criterion-referenced assessment systems developed, particularly since the late 1990s, which afforded teachers and students much more opportunity to practise for tests and improve coursework and modular grades through coaching, specific feedback, resubmission of work and re-sits of modular tests.

Thus, the evidence in England suggests that 'teaching to the test', drafting and redrafting coursework, and re-sitting tests are the most significant recent explanation for rising scores which tail off as teachers and students come to be about as efficient as they can be at scoring well on the tests within a regime of constant coaching and practice. Many research studies have reported an increasing focus on test preparation, and a wide range of international research evidence has also found that high-stakes testing leads to 'teaching to the test', including drafting and improving coursework and project work for final submission (Gillborn and Youdell, 2006; Hamilton *et al*., 2007; Klein *et al*., 2000; Linn, 2000; McNess *et al*., 2001; Torrance, 2007). In England, there has also been evidence of cheating, such is the pressure to report constantly improving pass rates, with teachers allowing more time in tests than they should when invigilating, indicating incorrect answers as they walk round by suggesting that students 'think again', and even changing scripts after collection, but before sending away for external marking (Daily Telegraph, 2016; Guardian, 2017; ITV, 2015; Mansell, 2015). Cheating is not unique to England of course. It is an increasingly visible effect of high-stakes accountability testing internationally (Nichols and Berliner, 2007; Stanford, 2013; Strauss, 2015). Overall then, we reached a situation in England around 2010 where scores and grades were continuing to rise but the validity, reliability and credibility of the standards achieved became subject to increasing doubt, and the educational experience of even the most successful students, let alone those who are not successful, was compromised.

6. One threat to validity leads to another...

National Curriculum Assessment

Identifying a problem is one thing of course, assuming that eliminating coursework and teacher involvement in assessment is the only solution is quite another. The research evidence indicates that it was the compounding effects of the accountability system, and the pressure to raise results at almost any (educational) cost that was and remains the real issue. Nevertheless, the Conservative government has moved swiftly to abolish coursework and other forms of school-based assessment of students by teachers. It has restricted National Curriculum Assessment at age 11 to externally set and marked tests of Maths, Reading and a Grammar, Spelling and Punctuation test, along with an internal teacher assessment of Writing. Science testing was ended in 2010. Biennial science tests are now used with a sample of students in an attempt to monitor standards in science over time. It is also important to note that the holistic construct of 'English' as a subject has effectively disappeared from National Curriculum Assessment – deconstructed into Reading, and Grammar, Spelling and Punctuation. Moreover, Reading is more of a written comprehension test than an assessment of reading as an activity (see: https://www.gov.uk/government/publications/key-stage-2-english-reading-test-framework). Writing is assessed internally but not

given the same prominence as Spelling, Grammar and Punctuation, while oral work – 'speaking and listening' – is no longer included as an integral element of English.

Changes to the National Curriculum itself are also being made, with the removal of 'levels' and 'level descriptors' (level 2, level 4, etc.) and the introduction of 'performance descriptors'. Thus, students will now be assessed at age 11 as 'Working at, above, or below National Standard' or to have achieved 'Mastery' – with the Standard being described in terms of curriculum content (DfE, 2014; STA, 2016). This is interpreted by some observers as a move towards mastering content and assessing greater depth of knowledge (Blatchford, 2015; Cambridge Primary Review, 2014)

These changes beg issues of validity at the level of individual measurement and evaluative utility. Testing at age 11 is now conducted almost exclusively via externally set and marked tests. The Standards and Testing Agency, an executive agency of the government's Department for Education, sets the tests and issues contracts to mark them – the current contract being held by Pearson (STA, 2016b). English in particular is now construed as a fragmented collection of constituent parts. Meanwhile, as regards consequential validity debates about the quality and accountability of individual schools and the English primary school system as a whole are based on a narrow set of tests in two subject areas.

7. One threat to validity leads to another...

GCSE

In parallel with insisting that tests of grammar, spelling and punctuation be instituted at age 11, GCSE has also come in for similar scrutiny. In terms very reminiscent of John Major's speech 20 years earlier, a 2010 Education White Paper stated:

> ...exams [will be] typically taken only at the end of the course... mark schemes [will include] spelling, punctuation and grammar... (DfE, 2010, p. 49)

The examination system in England is currently controlled by a regulatory body known as 'Ofqual' (Office of Qualifications and Examinations Regulation) and we can see in subsequent policy development how political intervention interacts with the technology of examinations. Criticisms of possible grade inflation and the reliability or otherwise of teacher assessment led initially to the introduction of 'controlled assessment' whereby school-based tasks were externally specified in greater detail and taken under controlled conditions in the classroom. This led to further problems of teacher workload and practising for the controlled assessments in advance of them actually being conducted (Baird *et al.*, 2013; Ofqual, 2013). As recently as 2013, Ofqual actually rejected a move back to wholly written examinations:

> There were suggestions that controlled assessment should be removed, and replaced with written exams. We did not see this as a viable option – many GCSEs include practical elements that cannot be assessed in a written exam. (Ofqual, 2013, p. 3)

However, further restrictions on the use of controlled assessment were introduced, which were then overtaken by government policy to end teacher assessment altogether. In a subsequent 'policy steer letter' to Ofqual, as they developed the details of the new system, the then Secretary of State for Education Michael Gove insisted that:

> The [new] qualifications should be linear, with all assessments taken at the end of the course… (Gove, 2013, p. 3)

In effect the GCSE system is being returned to an externally set and marked series of knowledge-based examinations. These proposals have met with significant educational challenge, including from key subject associations arguing for the importance of practical work in science, oral work in English, fieldwork in geography and so forth. Thus, for example, two core funders of curriculum development work in maths and science education, the Gatsby Foundation and the Wellcome Trust, produced a 'Policy Note' arguing that:

> Experiments are the essence of science and studying science without experimental work is like studying literature without reading books… No GCSE or A level science qualification should be awarded without evidence that students have developed hands on practical science skills. (Gatsby and Wellcome, 2013, p. 1)

Similar concerns were reported by the Nuffield Foundation with respect to parallel changes being enacted at A-level. Nuffield noted that modularisation had led to significantly increased numbers of students continuing with Maths, post-16, to A-level, especially amongst girls:

> A-level's modular structure has facilitated the growth in take-up over the past eight years, partly because students have been able to build qualifications and their own confidence, module by module. (Hillman, 2014, p. 11)

These are precisely the sorts of arguments that led to the expansion of coursework, modular work and practical work in the first place. However, these arguments have had very little impact on the changes that are being implemented. While practical activities are included in the curriculum, they are not included in the testing framework.

The first aspects of the reforms are being phased in and the implementation process through to 2017 is summarised in a 2014 Ofqual policy paper. Thus, for example, there will be no 'non-exam assessment' in GCSE Maths or English from 2015. In English, 'speaking skills will be assessed and reported as a separate grade' (Ofqual, 2014, p. 7) but this means they will not actually contribute to the English GCSE grade; when reported separately, they will become largely irrelevant. There will be no coursework or fieldwork in

Geography or History, though 'spelling, punctuation and grammar' will be allocated 5% of the marks in the exams in both subjects (p. 7). Spelling, punctuation and grammar will be allocated 20% of the marks in English (p. 7). In Geography, 'Schools must confirm students have carried out two pieces of fieldwork' (p. 7), but again, this will not contribute towards their grade, and two pieces of fieldwork in a 2-year course will hardly impact on the curriculum and student learning.

Arguments in Science continued until very recently. The pattern established for A-level Science is that 'practical skills will be recognised through compulsory practical work for which students will receive a "pass" or "fail" outcome, separate from their grade for the written exams' (Ofqual, 2014, p. 5). This means that practical laboratory work will not actually contribute to the reported subject grade and, contrary to Gatsby and Wellcome's insistence above, A-level Science qualifications will indeed be awarded without 'evidence that students have developed hands on practical science skills'. The evidence of a separate practical outcome will not actually contribute to the grade on the certificate. It may even be the case that students attain a pass grade in the written paper but a 'fail' in the separate practical work element; in such circumstances they will still be awarded a pass grade in Science. This situation has been described by the President of the UK Association of Science Education as a:

> highly dangerous experiment by Ofqual and ministers to separate the grade for assessed practical work from the main grades at A-level and GCSE. (Bell, 2015, p. 5)

The response of Ofqual has been to insist that: 'well-written questions can appropriately test candidates' knowledge of scientific experimentation' (Stacey, 2015, p. 2), which is hardly the same thing as hands-on practical competence. This is now the position with respect to GCSE Science as well:

> science GCSE qualifications, to be taught from September 2016, will assess practical work using written exam questions in place of controlled assessment... Students' knowledge of the practical work... will be assessed in exam questions at the end of the course. (Ofqual, 2015, p. 1)

'Practical activities' will still feature in GCSE Science courses but they will not form part of the assessment for grading purposes:

> each exam board will specify a minimum number of practical activities that students must complete, set no lower than 8 for individual sciences and 16 for combined science. (Ofqual, 2015, p. 1)

Interestingly, the position of Ofqual is grounded not just in Conservative government concerns about grade inflation, but also in concerns about the educational experience of teachers and students. Glenys Stacey, the then Chief Executive of Ofqual, in a speech to the Association of College Examinations Officers argued that

> for too long the assessment system has encouraged… teachers to repeat and rehearse the same narrow group of practicals in order to achieve the best possible grades… assessment drives and trammels what is taught. (Stacey, 2015, pp. 1–2)

Again, however, it is not clear how simply abandoning the integration of assessment of practical laboratory work with final examinations will improve this experience or the acquisition of practical skills. Rather, teachers will simply concentrate on making sure that students can answer questions about conducting experiments. The issue is the pressure of accountability and the pursuit of higher grades at any cost, not the method of assessment.

8. Where do we go from here? Different versions of what is meant by validity, credibility and 'educational standards'

So, final examinations are set to dominate all aspects of English education. Where does this leave us with respect to debates about validity, a wider curriculum, new forms of assessment and the improvement of educational standards? There are very good reasons for developing new forms of assessment – reasons which derive from what we might term an examining perspective as well as from an educational perspective. Educationally, new curriculum content which looks to develop knowledge and skills of investigation, data analysis, report-writing, team-working and so forth, clearly looks to assessment methods which can both identify and report such outcomes, and underpin their development. As Resnick and Resnick (1992) put it more than 20 years ago:

> You get what you assess; you don't get what you don't assess;
> you should build assessment towards what you want…to teach… (p. 59)

Cisco, Intel and Microsoft developed the 'Assessment and Teaching of 21st Century Skills' project (ATCS, 2010) which lists '10 skills' essential to economic and social success in the twenty first century including: creativity and innovation, critical thinking, problem-solving, decision-making, learning to learn, metacognition, communication, collaboration (teamwork), citizenship – local and global, personal and social responsibility – including cultural awareness and competence (ATCS, 2010, pp. 1–2). ATCS argues that not only should such 'soft skills' be developed by education systems but also it is important that assessment procedures and practices should change to incorporate and encourage their development:

> The crux of 21st century skills is the need to integrate, synthesize and creatively apply knowledge in novel situations… 21st century assessments must systematically ask students to apply… knowledge to critical thinking, problems solving and analytic tasks… (2010, p. 6)

Such arguments echo those of Wellcome and Nuffield with respect to the development of practical and analytic skills in science education, noted earlier.

From an examining perspective, new curriculum goals demand new forms of assessment in order to report grades with validity and reliability – coursework, fieldwork, oral work and so forth can capture different outcomes from end-of-course written tests, including ephemeral outcomes such as confidence in discussion and problem-solving. When we also then add in ideas about formative assessment and changes in pedagogy, including students drafting work, receiving feedback on it and redrafting it for final submission, we produce a potentially very positive situation in which students can be supported to develop their knowledge and understanding of subject matter over time and produce their best possible work for examination purposes. In principle, this should constitute the core of any attempt to broaden and raise 'educational standards'.

Equally, however, such arguments present the traditional demands of an assessment system, with very great challenges. A further paper in the ATCS series argues that:

> Problem solving assessment tasks will need to represent... open-ended tasks permitting multiple appropriate methods for eliciting evidence of how well learners plan, conduct and interpret evidence... [but]... the state of practice for assessing 21st century skills integrated into learning activities remains in its infancy... (Scardamalia *et al*., 2010, p. 31)

As the above quote implies, and some of this paper's earlier discussion noted, much current empirical evidence suggests that it is very difficult to produce reliable results from these sorts of extended 'authentic' or 'performance' assessments (Baker and O'Neil, 1994; Johnson, 2013; Koretz, 1998; Stanley *et al*., 2009).

Enacted in a context of intense accountability pressures, as has been the case in England, flexible, formative assessment practices can lead to little more than coaching students to meet examination criteria thus undermining the validity and credibility of results. In previous research, I have identified this as 'criteria compliance' (Torrance, 2007, p. 282). It also leaves students with little in the way of in-depth understanding, and expecting to be similarly coached at university and indeed in employment. A key issue for this paper, however, is that improving the validity and reliability of open forms of assessment involving teacher assessment of their students in school is possible, and it is with such issues that assessment policy, research and development should be trying to engage (see also CISCO, 2010; Pearson, 2016; Scardamalia *et al*., 2012).

From a political perspective, however, evidence of grade inflation has been seized upon to turn the educational clock back, once more, in the name of protecting traditional educational standards. A major issue is the likely impact on the curriculum and on the educational experience and outcomes of future cohorts of students. As a recent OECD (2013) report noted:

> Because of their role in... accountability, evaluation and assessment systems can distort how and what students are taught... It is important to minimise these

> unwanted side-effects… Thus performance measures should be broad… drawing on quantitative and qualitative data… (pp. 2–3).

In England, we are faced with a number of apparently irreconcilable economic, social and political pressures. Neo-liberal human resource development theory produces policy which seeks to extend educational provision to the widest possible student group. Pursuing different curriculum goals requires the development and use of a wider range of assessment methods. In tandem with these developments, educational arguments seek to implement formative feedback to promote learning. However, accountability measures introduced to reassure government that it is receiving an appropriate return on its investment in education exert pressure on schools and teachers to raise grades, irrespective of whether or not they reflect real improvements in student achievement. The validity of the assessment is thus called into question and the resultant rise in examination passes, supposedly the key indicator of rising educational standards and the success of policy, can be interpreted as evidence of falling standards.

We are left with some acute questions: How can the assessment of a wide range of individual achievements be reconciled with political pressure to use examination results to measure the system as a system? How can flexible assessments be developed that respond to the need for a more expansive and flexible curriculum without compromising the quality of the educational challenge that such a curriculum should comprise? and How is it that politicians can set aside decades of research into how to construct valid and reliable assessments, especially when, in so doing, they undermine their own (apparent) aspirations to develop more flexible skills and abilities supposedly required for the ‘knowledge economy’? With respect to this last question at least part of the answer must reside in the lack of institutional memory in educational policy-making (and indeed in some parts of the assessment research community) such that disconnected policy initiatives and interventions stimulate similarly disconnected ‘bursts’ of research activity, without any sustained evidence-based interaction between researchers and policymakers. Developing sound approaches to authentic or performance assessment should be the focus of policy and research but at present this is not the case.

It is clear that for purposes of educational quality and accountability, educational systems should employ multiple measures of achievement if possible, but equally it is clear that the greater the scale and scope of the assessment system, the more difficult it is to accommodate different methods of assessment at a high level of quality. Glenys Stacey, in her 2015 speech, also noted that ‘more than 22 million exam scripts and pieces of coursework were marked last year’ (Stacey, 2015, p. 6). And these figures relate only to GCSE and A-level. A further 3.8 million scripts were set and marked for National Curriculum Assessment at age 11 in 2015 (STA, 2016b). Similarly, the more individual student achievement is tied to system accountability, the more accountability measures will dominate pedagogy and student experience. Therefore policy should

(1) decouple accountability measures from routine student assessment and address the monitoring of standards over time by use of specifically designed tests with small national samples;
(2) re-conceptualise the development of educational standards by starting from the perspective of the curriculum: i.e. put resources and support into re-thinking curriculum goals for the twenty-first century and developing illustrative examples of high-quality assessment tasks that underpin and reinforce these goals, for teachers to use and adapt as appropriate.

The introduction of the new national reference test in 2017 may be interpreted as something of a move in the direction of (1) above. A small sample of students (c. 18,000) nationally will take tests in English and Maths, just prior to sitting GCSEs in the summer, the intention being to monitor standards of English and Maths, over time, independent of GCSE (NFER, 2017; Ofqual, 2015). At present, however, the tests are intended to contextualise GCSE results and help to evaluate any further debate about grade inflation (or indeed the reverse, given the arguments in this paper that headline results may now continue to fall). The pressure of accountability, ensconced in 'Progress 8' and the new GCSE grading system (with '5' being the lowest acceptable 'pass' grade), remains.

Similarly, as noted above, there is significant international interest in developing new forms of assessment which would underpin the policy developments outlined in (2) above. There is also considerable research evidence about how models of moderation could be integrated with and underpin continuing professional development for teachers involved in assessment such that issues of validity and reliability of results could be addressed (Black *et al*., 2010, 2011; Stanley *et al*., 2009). The issue is one of political will and where best to invest in the education and assessment of students.

We are a long way from the scenario envisaged in (1) and (2) in England at present. Indeed, we are making every mistake that the recent OECD (2013) report warned against. Equally, however, the OECD itself might pay more attention to the empirical evidence reported above. The policy aspiration of OECD, that the same assessment 'tools' can both measure 'how well students are learning' and provide information to 'society at large about educational performance' (OECD, 2013, p. 1), is very misplaced. When the same assessments are used to measure individual achievement and the effectiveness of the system via processes of accountability, the pressure to improve grades undermines belief in their credibility. Thus when, for whatever reasons, more and more students pass more and more exams, educational arguments about the wider purposes of assessment are all too easily set aside.

9. ACKNOWLEDGEMENTS

Earlier versions of this paper were presented as 'Validity, or the lack of it, in English National Curriculum Assessment and School Examinations' to the annual

conference of the British Educational Research Association (BERA) 13-15 September 2016, Leeds, UK, and the Australian Association for Research in Education (AARE), Fremantle, 29 November–3 December 2015. The author is grateful to Professor Val Klenowski for organising the symposia in which the papers were presented and to Queensland University of Technology and the Australian Association for Research in Education for their support.

10. Disclosure statement

No potential conflict of interest was reported by the author.

11. Notes

[1] That is, the equivalent of five GCSEs at grades A*–C: the top GCSE grades of A*–C are officially accepted as the equivalent of the old O-level passes; the percentage of students gaining at least five A*–Cs is the commonly accepted and reported measure of a good secondary education; the percentage of students gaining at least five A*–Gs (the full range of grades) is the commonly accepted and reported measure of a minimally satisfactory secondary education.

[2] Not every year's results are recorded in Figures 1 and 2; rather, sufficient years are recorded to indicate trends over time along with key dates indicating changes of government and/or which government has variously used and dropped as indicators of progress. Also, overall pass rates conceal other issues. Within these general trends different sub-groups perform better than others, and results vary by social class, gender and race. Students of Chinese origin do best, and white British working boys and Black Caribbean working class boys do worst (DfE, 2016b). For the purposes of this paper, however, the point is that pass rates have been rising since the 1970s. Pass rates cited derive from Torrance (2011) for 1975–2010 and DfE (2016b) for pass rates since 2010.

12. References

Assessment and Teaching of 21st Century Skills (ATCS). (2010) *Draft White Paper 1: Defining 21st Century Skills*. Available at: http://atc21s.org/wp-content/uploads/2011/11/1-Defining-21st-Century-Skills.pdf

Baird, J., Ahmed, A., Hopfenbeck, T. N., Brown, C. and Elliott, V. (2013) *Research evidence relating to proposals for reform of the GCSE*. OUCEA Report. Oxford, Oxford University.

Baird, J. and Black, P. (2013) Test theories, educational priorities and reliability of public examinations in England, *Assessment in Education*, 28 (1), 5–21.

Baker, E. and O'Neil, H. (1994) Performance assessment and equity: a view from the USA, *Assessment in Education*, 1 (1), 11–26. doi:10.1080/0969594940010102

Bell, D, (2015) 'Science Education: Trusting the Front Line' Keynote speech to Association of Science Education Annual Conference, Reading.

Beloe Report. (1960) *Secondary school examinations other than the GCE*. Ministry of Education HMSO. London, HMSO.

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2010) Validity in teachers' summative assessment, *Assessment in Education*, 17 (2), 215–232. doi:10.1080/09695941003696016

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2011) Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education*, 18 (4), 451–469. doi:10.1080/0969594X.2011.557020

Blatchford, R. (2015) Differentiation is Out. Mastery is the New Classroom Buzzword, *The Guardian*. Available at: https://www.theguardian.com/teacher-network/2015/oct/01/mastery-differentiation-new-classroom-buzzword

Broadfoot, P. (1996) *Education, Assessment and Society* (Maidenhead, Open University Press).

Brown, P. and Lauder, H. (Eds) (1992) *Education for Economic Survival* (London, Routledge).

Cambridge Primary Review. (2010) Available at: http://cprtrust.org.uk/

Cambridge Primary Review. (2014) *From Levels to Performance Descriptors: Labelling by Another Name?* Available at: http://cprtrust.org.uk/cprt-blog/performance-descriptors-labelling-by-another-name/

CISCO. (2010) *Assessment and Teaching of 21st century skills*. Available at: http://www.cisco.com/c/dam/en_us/about/citizenship/socioeconomic/docs/ATC21S_Exec_Summary.pdf

Cohen, L. and Deale, R. (1977) *Assessment by Teachers in Examinations at 16+* Schools Council Examinations Bulletin 37 (London, Evans/Methuen).

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. and York, R. L. (1966) *Equality of Educational Opportunity*(Washington, DC, US Department of Health Education and Welfare).

Daily Telegraph. (2016) *Boarding School Principal 'Gave Go-Ahead on Exam Cheating'*. Available at: http://www.telegraph.co.uk/education/2016/10/10/boarding-school-principal-gave-go-ahead-on-exam-cheating/

Daugherty, R. (1995) *National Curriculum Assessment: A Review of Policy 1987–1994* (London,Routledge).

Dearing, R. (1994) *The National Curriculum and Its Assessment: Final Report* (London, School Curriculum and Assessment Authority).

Department for Education (DfE). (2010) *The Importance of Teaching*, Cmnd 7980.

Department for Education (DfE). (2014) *Performance Descriptors*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/368298/KS1-KS2_Performance_descriptors_consultation.pdf

Department for Education (DfE). (2016a) *Primary School Tests Show Schools Rising to the Challenge*. Available at: https://www.gov.uk/government/news/new-primary-school-tests-show-schools-rising-to-the-challenge

Department for Education (DfE) (2016b) *Revised GCSE and Equivalent Results in England 2014 to 2015*. SFR 01/2016. Available at: https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2014-to-2015

Department for Education and Science (DES). (1982) *Examinations at 16 Plus: A Statement of Policy* (London, HMSO).

Department of Education and Science. (1988) *Task group on assessment and testing*. The TGAT Report (London, DES).

DfE. (2017) *Progress 8 and Attainment 8*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/583857/Progress_8_school_performance_measure_Jan_17.pdf

Gatsby Foundation and Wellcome Trust. (2013) *Policy Note: Assessment of Practice Work in Science April 2013*. Available at: http://www.gatsby.org.uk/uploads/education/reports/pdf/practical-science-policy-note.pdf

Gibb, N. (2015, October 17) The Beginning of Inflation-Proof Excellence in Schools, *Daily Telegraph*. Available at: http://www.telegraph.co.uk/education/educationopinion/11937788/The-beginning-of-inflation-proof-excellence-in-schools.html

Gillborn, D. and Youdell, D. (2006) Educational triage and the D-to-C conversion: suitable case for treatment? In H. Lauder, P. Brown, J. Dillabough and A. H. Halsey (Eds) *Education, Globalisation and Social Change* (Oxford, Oxford University Press).

Gipps, C. and Murphy, P. (1994) *A Fair Test? Assessment, Achievement and Equity* (Maidenhead, Open University Press).

Gipps, C. and Stobart, G. (2009) Fairness in assessment. In J. Wyatt-Smith and J. Cummings (Eds) *Educational Assessment in the 21st Century* (Dordrecht, Springer).

Gove, M. (2013, February 6) *Ofqual Policy Steer Letter: Reforming Key Stage 4 Qualifications*, DfE, Sanctuary Buildings.

Greening, J. (2017, March 28) *Letter to Chair of House of Commons Education Select Committee*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/603594/ESC_letter.pdf

Guardian. (2015) *Ministers Consider National Tests for 7 Year Olds*. Available at: https://www.theguardian.com/politics/2015/nov/01/nicky-morgan-national-tests-primary-school-england

Guardian (2017) *Subject: Staff at 'Outstanding' London School Suspended Over Alleged Exam Cheating*. Available at: https://www.theguardian.com/education/2017/feb/10/staff-at-outstanding-london-school-suspended-over-alleged-exam-cheating?CMP=share_btn_link

Hamilton, L., Stecher, B., March, J., McComb, J., Robyn, A., Russell J., Naftel S. and Barney H. (2007) *Standards-Based Accountability under No Child Left Behind* (Santa Monica, Rand Education).

Hargreaves, A. (1989) *Curriculum and Assessment Reform* (Maidenhead, Open University Press).

Hargreaves, A. (1994) *Changing Teachers, Changing Times* (London, Routledge).

Harlen, W. (2005) Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes, *Research Papers in Education*, 20 (3), 245–270. doi:10.1080/02671520500193744

Hillman, J. (2014) *Mathematics After 16: The State of Play, Challenges and Ways Ahead* (London, Nuffield Foundation).

Horton, T. (Ed.) (1988) *GCSE: Examining the New System* (London, Harper and Row).

Hoste, R. and Bloomfield, B. (1975) *Continuous Assessment in the CSE* Schools Council Bulletin 31 (London, Evans / Methuen).

House of Commons Select Committee for Children, Schools and Families. (2008, May 13) *Select Committee Reported on BBC*. Available at: http://news.bbc.co.uk/1/hi/education/7396623.stm

Isaacs, T. (2010) Educational assessment in England, *Assessment in Education*, 17 (3), 315–334. doi:10.1080/0969594X.2010.491787

ITV. (2015) *Exposure: Making the Grade, Documentary Investigating Allegations of Cheating in UK GCSE Examinations*. Available at: https://www.itv.com/itvplayer/exposure/series-27/episode-1-exposure-making-the-grade

Johnson, S. (2013) On the reliability of high-stakes teacher assessment, *Research Papers in Education*, 28 (1), 91–105. doi:10.1080/02671522.2012.754229

Joseph, K. (1986) The role and responsibility of the secretary of state. In *DES Better Schools: Evaluation and Appraisal Conference* (London, HMSO).

Klein, S., Hamilton, L., McCaffery, D. and Stecher B. (2000) What do test scores in Texas tell us? *Education Policy Analysis Archives*, 8, 49. doi:10.14507/epaa.v8n49.2000

Koretz, D. (1998) Large-scale portfolio assessments in the US: evidence pertaining to the quality of measurement, *Assessment in Education*, 5 (3), 309–334. doi:10.1080/0969595980050302

Lauder, H., Brown, P., Dillabough, J.-A. and Halsey, A. H. (Eds) (2006) *Education, Globalisation and Social Change* (Oxford, Oxford University Press).

Linn, R. (2000) Assessments and accountability, *Educational Researcher*, 29, 4–16.

Mansell, W. (2015) Exam Cheating: Time for Full Disclosure, *NAHT*. Available at: http://www.naht.org.uk/welcome/news-and-media/blogs/warwick-mansell/exam-cheating-time-for-full-disclosure/

McNess, E., Triggs, P., Broadfoot, P., Osborn, M. and Pollard, A. (2001) The changing nature of assessment in English primary schools: findings from the PACE project 1989-1997, *Education 3-13*, 29 (3), 9–16. doi:10.1080/03004270185200301

Messick, S. (1989) Validity. In R. Linn (Ed) *Educational Measurement* (3rd edn) (New York, MacMillan).

Morgan, N. (2015a) One Nation Education. *Speech to Policy Exchange, London*. Available at: https://www.gov.uk/government/speeches/nicky-morgan-one-nation-education

Morgan, N. (2015b) New Reforms to Improve Standards and Improve Behaviour, *Press Release*. Available at: https://www.gov.uk/government/news/new-reforms-to-raise-standards-and-improve-behaviour

Newton, P. and Shaw, S. (2016) Disagreement over the best way to use the word 'validity' and options for reaching consensus, *Assessment in Education*, 23 (2), 178–197. doi:10.1080/0969594X.2015.1037241

NFER. (2017) *National Reference Test*. Available at: https://www.nfer.ac.uk/schools/national-reference-test/

Nichols, S. and Berliner, D. (2007) *Collateral Damage: How High Stakes Testing Corrupts America's Schools Harvard* (Cambridge, MA, Education Press).

Norwood Report. (1943) *Curriculum and Examinations in Secondary Schools* (London, HMSO).

NUT. (2016) *Justine Greening Announcement on Primary Assessment*. Available at: https://www.teachers.org.uk/news-events/press-releases-england/justine-greening-announcement-primary-assessment

OECD. (2013) *Synergies for Better Learning: An International Perspective on Evaluation and Assessment* (Paris, OECD).

Office for Standards in Education. (2006) *The Annual Report of Her Majesty's Chief Inspector of Schools 2005/06* (London, OfSTED). Available at: http://www.ofsted.gov.uk

Ofqual. (2013) *Review of Controlled Assessment in GCSEs*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/377903/2013-06-11-review-of-controlled-assessment-in-GCSEs.pdf

Ofqual. (2014) *An Update on the Reforms Being Made to GCSEs*, Ofqual/14/5404

Ofqual. (2015) *Consultation Outcome: Assessing Practical Work in GCSE Science*. Available at: https://www.gov.uk/government/consultations/assessing-practical-work-in-gcse-science

Ofqual. (2016a) *An Investigation into the 'Sawtooth Effect' in GCSE and AS/A Level Assessments*, Ofqual/16/6098.

Ofqual. (2016b) *Some Thoughts on Comparable Outcomes*. Available at: https://ofqual.blog.gov.uk/2016/06/03/some-thoughts-on-comparable-outcomes/

Pearson. (2016) *Edexcel International GCSEs*. Available at: http://qualifications.pearson.com/content/dam/pdf/International%20GCSE/General/IG-Transferable-Skills-Information-Pack.pdf

Perry, E. and Francis, B. (2010) *The Social Class Gap for Educational Achievement: A Review of the Literature RSA*. Available at: https://www.thersa.org/globalassets/pdfs/reports/rsa-social-justice-paper.pdf

Popham, J. (1997) Consequential validity: right concern, wrong concept, *Educational Measurement: issues and Practice*, 16 (2), 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Porter, A. (2008) 'Ed Balls Signals End to Sats Exam from Next Year, *Daily Telegraph*. Available at: http://www.telegraph.co.uk/news/2699916/Ed-Balls-signals-end-to-Sats-exam-from-next-year.html

Resnick, L. and Resnick, D. (1992) Assessing the thinking curriculum. In B. Gifford and M. O'Connor (Eds) *Future Assessments: Changing Views of Aptitude, Achievement and Instruction* (Boston, MA, Kluwer).

Scardamalia, M., Bransford, J., Kozma, B. and Quellmalz, E. (2010) *ATCS Draft White Paper 4: New Assessments and Environments for Knowledge Building*. Available at: http://atc21s.org/wp-content/uploads/2011/11/4-Environments.pdf (accessed 8 August 2012).

Scardamalia, M., Bransford, J., Kozma, R. and Quellmalz, E. (2012) New assessments and environments for knowledge building. In P. Griffin, B. McGaw and E. Care (Eds) *Assessment and Learning of 21st Century Skills* (Dordrecht, Springer Science +Business Media B.V), 231–300.

Serin, S. (2005) Socio-economic status and academic achievement, *Review of Educational Research*, 75 (3), 417–453. doi:10.3102/00346543075003417

STA. (2016) *Interim Teacher Assessment Frameworks*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538415/2017_interim_teacher_assessment_frameworks_at_the_end_of_key_stage_2_150716_PDFa.pdf

Stacey, G, (2015) 'Tomorrow's World' Speech to Association of College Examinations Officers' Conference

Standards and Testing Agency (STA). (2016a) *Information for Parents*: 2016 *National Curriculum Tests at the End of Key Stages 1 and 2 STA/16/7562/Pke*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/520553/Information_for_parents_-_2016_NCTs_at_the_end_of_key_stages_1_and_2_27042016__2__PDFA.pdf

Standards and Testing Agency (STA). (2016b) *Annual Report and Accounts for the year ended* 31 March 2016 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538564/STA_Annual_report_and_accounts.PDF

Stanford, J. (2013) High-Stakes Testing Makes Cheating Inevitable, *The Huffington Post*. Available at: http://www.huffingtonpost.com/jason-stanford/standardized-test-cheating_b_2993239.html see also: http://www.huffingtonpost.com/news/standardized-test-cheating/

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009) *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development* (Oxford, OUCEA).

Stobart, G. (2001) The validity of National Curriculum, Assessment, *British Journal of Educational Studies*, 49 (1), 26–39. doi:10.1111/1467-8527.t01-1-00161

Stobart, G. (2009) Determining validity in National Curriculum Assessments, *Educational Research*, 51 (2), 161–179. doi:10.1080/00131880902891305

Strauss, V. (2015) How and Why Convicted Atlanta Teachers Cheated on Standardized Tests, *The Washington Post*. Available at: https://www.washingtonpost.com/news/answer-sheet/wp/2015/04/01/how-and-why-convicted-atlanta-teachers-cheated-on-standardized-tests/?utm_term=.d6e8c4331ebe

Torrance, H. (2007) Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning, *Assessment in Education*, 14 (3), 281–294. doi:10.1080/09695940701591867

Torrance, H. (2011) Using assessment to drive the reform of schooling: time to stop pursuing the chimera? *British Journal of Educational Studies*, 59 (4), 459–485. doi:10.1080/00071005.2011.620944

Torrance, H. (2016) Blaming the victim: assessment, examinations, and the responsibilisation of students and teachers in neo-liberal governance, *Discourse: studies in the Cultural Politics of Education*. doi:10.1080/01596306.2015.1104854

Torrance, H. (2009) Using assessment in education reform: policy, practice and future possibilities. In H. Daniels, H. Lauder and J. Porter (Eds) *Knowledge, Values and Educational Policy* (London, Routledge).

Torrance, H. (2003) Assessment of the National Curriculum in England. In T. Kellaghan and D. Stufflebeam (Eds) *International Handbook of Educational Evaluation* (Dordrecht, Kluwer).

Turner, D. (2009) Fresh Blow for Sats as Science Test Scrapped, *Financial Times*. Available at: https://www.ft.com/content/02301ef4-3b5e-11de-ba91-00144feabdc0

Waddell Report. (1978) *School Examinations* (London, HMSO).

Weiner, K. (2013) Consequential validity and the transformation of tests from measurement tools to policy tools, *Teachers College Record*, 115 (9), 1–6.

Wolf, A. (1995) *Competence-Based Assessment* (Maidenhead, Open University Press).

Correspondence
Harry Torrance
Education and Social Research Institute
Manchester Metropolitan University
Brooks Building
Bonsall Street
Manchester M15 6GX
Email: h.torrance@mmu.ac.uk