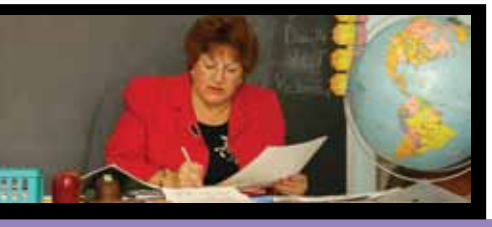
Assessment in schools
Fit for purpose?

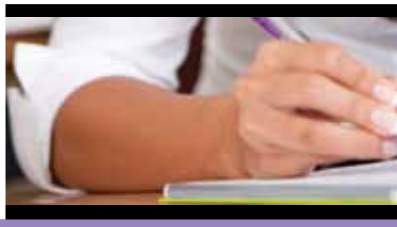
A Commentary by the Teaching and Learning Research Programme

ARG
assessment reform group

T·L·R·P
TEACHING & LEARNING RESEARCH PROGRAMME

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL





Education is recognised across the world as perhaps the most vital public service of all. But how can we measure its effect? Assessment is essential to allow individuals to get the educational support they need to succeed, to see the effectiveness of different educational methods, and to ensure that education budgets are being spent effectively. Inevitably, assessment also risks marking teachers, learners and institutions as successes or failures.

This Commentary on assessment is a joint venture undertaken by the Teaching and Learning Research Programme and the Assessment Reform Group. The TLRP has been the largest ESRC programme, carrying out research into all aspects of education from preschool to adult and workplace learning. Many of its projects have considered aspects of assessment, especially its highly impactful work on Assessment for Learning.

The Assessment Reform Group has been in existence as an informal network for even longer than the TLRP. Its emphasis has been on ways in which assessment can help to advance learning rather than merely measuring it. This is expected to be the Group's last major publication and it is a pleasure to pay tribute to its contribution to advancing our ideas about appropriate ways of assessing educational outcomes.

Current developments in education, such as the rethinking of SATs as they are applied to schoolchildren in England, suggest that the ARG's views on the right and wrong way to assess educational progress are having steadily more influence, and that TLRP research has helped to build the evidence needed to improve assessment, and the educational outcomes which it is meant to underpin.

This is the thirteenth TLRP Commentary, and is intended like it predecessors to bring top-quality research to practitioners, policy-makers and the interested public. We hope you enjoy it, and welcome your feedback via our website www.esrc.ac.uk.

Professor Ian Diamond FBA AcSS
Chief Executive
The Economic and Social Research Council

contents

Introduction: the policy context	4
Purposes of assessment	7
Quality	9
Quality in formative assessment/assessment for learning	9
Quality in summative assessment	12
Quality in accountability	17
Four pressing challenges for policy-makers	20
Putting effective in-class assessment into practice system-wide	20
Enhancing confidence in tests and examinations	24
Justifying the costs of assessment	26
Avoiding micro-management	28
Conclusion: enhancing public understanding for wise decision-making	30
About Warwick Mansell, Mary James and the Assessment Reform Group	31

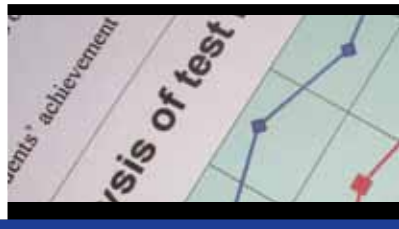
Warwick Mansell

Mary James

Assessment Reform Group

This commentary has been written by Warwick Mansell, Mary James and the Assessment Reform Group on behalf of TLRP (July 2009). For further information see inside of back cover.

This publication should be cited as: Mansell, W., James, M. & the Assessment Reform Group (2009) *Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme.* London: Economic and Social Research Council, Teaching and Learning Research Programme.



Introduction: the policy context

Perhaps no area of education policy is as contentious – or as consistently newsworthy – as assessment. Recent headlines show how emotive and controversial it can be: "Tests blamed for blighting children's lives"; "New fears over dumbing down of key exams"; "Science exam standards 'eroded'"[1].

The public, formal, face of assessment – typically "high-stakes" examinations such as GCSEs, A-levels, Scottish Highers, the Welsh Baccalaureate or national curriculum tests in England – often dominates debate. But all good teachers also use assessment informally in the classroom to judge what progress pupils have made with their understanding, and to provide information on how they can be helped to move forward.

It is the interplay between the various forms of assessment, the uses to which the results of assessment judgments are put, and the consequent effects on teaching and learning that make this such a fascinating, significant and yet contentious, area for research, policy-making and discussion.

This commentary is an attempt to clarify some of the debates, bringing to bear the findings from extensive research, especially research from the Teaching and Learning Research Programme (TLRP) and the Assessment Reform Group (ARG), and to argue that this evidence needs to inform any future developments in assessment policy.

Many assessment debates are universal such as how assessment might best support learning and teaching, and how assessment is used to provide information on the progress of individuals, schools and countries. This commentary includes examples of how these play out in the four countries of the UK. However, in England, many of these questions have been particularly contentious and offer opportunities to explore tensions in some depth. The intention of this commentary is not to offer an overview of practice in each of the four countries within the UK but to use different experiences to illuminate more general concerns.

Not only is assessment an area of constant media scrutiny, it is also ever-changing. England's qualifications system is in the middle of the largest set of revisions it has undergone for decades, with the advent of reformed A-levels and GCSEs, the impending nationwide introduction of "functional skills" tests in English, mathematics and ICT and the recent creation of yet another set of work-related and general vocational courses. Scotland reformed its national qualification system in 2000 and further changes are being planned as part of the new Curriculum for Excellence, where there is a focus on skills, attributes and capacities for learning, life and work in the 21st century.

Radical changes to assessment up to the age of 14 are also being made in every part of the UK.

In Scotland, the development of a coherent assessment system, Assessment is for Learning, has been a government priority since 2001. There has been a major focus on assessment for learning – a concept that we discuss below. The Scottish Government no longer collects information on all pupils through national assessments but does monitor national achievement through the Scottish Survey of Achievement sample survey. Wales discontinued key stage testing in 2005 and, as well as prioritising assessment for learning, is now relying on the moderated judgments of teachers to sum up attainment at the end of key stages 2 and 3. Northern Ireland completed the termination of statutory testing at key stages 1-3 in 2006 and continues to promote assessment for learning and teacher assessment more generally in the revised arrangements for the new curriculum.

1 "Tests blamed for blighting children's lives", Guardian, 20/02/09; "New fears over dumbing down down of key exams" Observer 15/02/09; "Science exam standards 'eroded' ", BBC News Online, 26/11/08

In 2008, England's key stage 3 tests were scrapped suddenly after a marking failure that saw tens of thousands of pupils receive their results late. At key stage 2, a trial is investigating replacing the current end-of-key-stage tests with shorter ones that pupils could take at any time between the ages of seven and 11. However, the future of key stage 2 tests remains uncertain. Citing concern about evidence of "teaching to the test", the Conservatives have announced that they would scrap key stage 2 tests and replace them with tests to be taken on entry to secondary schools[2]. Proposed changes to the school accountability system in England, which will include the introduction of a new "report card" system of communicating judgments on school quality to parents, will also have implications for assessment, especially if results of a range of quite different measures are aggregated and reduced to a single grade.

Meanwhile, with less publicity, there have been attempts by the government in England to improve the quality of teacher assessment by initiatives called "Assessment for Learning" and "Assessing Pupils' Progress". A "chartered educational assessor" grade, providing recognition for those who have proved themselves excellent in this field, has also been developed. Finally, complex data analysis systems, centring on pupil test and examination results, have been designed with the aim both of informing the public on the quality of education available in individual schools and of assisting teachers to help pupils improve.

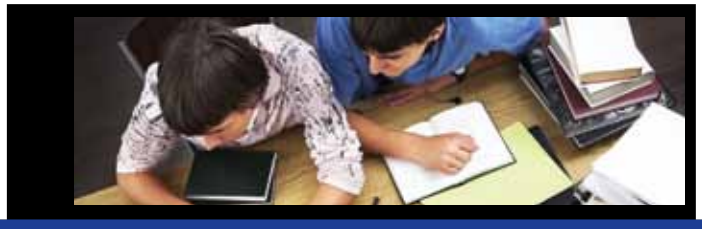
If policy change has been hyperactive, in one sense this is not surprising because assessment has been asked to perform an increasing number of functions in recent years: from judging individual pupils to evaluating schools and monitoring national performance. A key question for research has been whether it is effectively meeting all of the goals that have been set for it, and indeed, whether it is possible for it ever truly to fulfill so many aims simultaneously.

Why should this matter?

The nature and impact of assessment depends on the uses to which the results of that assessment are put. A system whose main priority is to generate information for internal use by teachers on the next steps in pupils' learning may have different characteristics and effects from one where the drive is to produce a qualification which will provide a grade on which an employer or a university admissions tutor might rely in order to judge the suitability of a candidate for employment or further study.

Particularly if results are to be used in "high stakes" situations, important questions must be asked. If, for example, the fate of a school may hang on a single set of test results, are the data they generate reliable enough to serve as a measure of the overall quality of that institution? Are they valid in measuring all that is felt to be important in education? Do they ultimately provide information that will help pupils improve? Finally, does the act of placing weight on the results of these tests affect the teaching that takes place in preparation for these assessments and, if there are negative impacts, are we prepared to tolerate them?

2 "Tories plan to scrap primary Sats", BBC News Online, 14/06/08



These are just some of the questions, and controversies, thrown up by this subject. As the number of purposes, and the weight placed on them, has grown, so these dilemmas become more pressing. For example, in 2008, the government in England placed a closure threat over any secondary school that failed to ensure that at least 30 per cent of its pupils achieved five GCSE A*-C grades including English and mathematics within three years [3]. Warnings have also been delivered to primary schools below a performance threshold[4]. How valid is this use of assessment data for making decisions on whether schools are fit to educate pupils?

Scotland adopts a different perspective on the use of assessment data to judge schools. For instance, it has maintained a survey approach to national monitoring, arguing that it offers the potential to provide information on pupil achievement on more of what matters across the curriculum; and that, because individual teachers and schools are not identified, teachers are not threatened by the survey and therefore are not tempted to narrow the curriculum or teach to the test. The decision in Scotland to stop collecting national assessment results for all pupils in all schools was taken because policy-makers recognised that teachers were teaching to the tests and that, although it appeared that results were improving, it was more likely to be that teachers were getting better at rehearsing children for the tests.

This commentary will argue that assessment must, in all cases, promote, rather than undermine, good education. Policy-makers need to keep the needs of pupils to the fore, and ensure that any evaluation of new developments in assessment is carried out with careful consideration of the consequences, both intended and unintended, for the quality of learning which results. In such a pivotal area of education, it is fortunate that a wealth of research is available.

Sources:

Whetton, C. (2009) A brief history of a testing time: national curriculum assessment in England 1989-2008, *Educational Research*, 51(2): 137-159.

3 "'No excuses' on school results", BBC News Online, 10/06/08

4 "Where it is clear that a school will not be able to improve their results enough to move above the floor target in 2010, even with additional support, a more radical solution such as closure or use of the local authority's statutory intervention powers will need to be considered." Government target-setting guidance for English primaries, sent to local authorities in October 2008.

Purposes of assessment

What are the purposes of the UK's assessment systems? This question has many answers.

From its close focus on helping teachers and pupils build a shared understanding of the progress the pupil has made in order to provide pointers for further development, to the wide-angle view of national education standards that test and examination results purport to provide, assessment information is now used in a multitude of ways. Yet the assessments on which these data rest have not been designed to support all such uses and interpretations. Whether they are capable of supporting the inferences drawn from them is a key question.

It is helpful to make a distinction, here, between the **intended use**, or uses, of assessment data, and their **actual uses**. Assessments are often designed quite differently to ensure their fitness for different purposes. Or, to put it another way, results that are fit to be used for one particular (intended) purpose may not be fit to be used for another, regardless of whether they are actually used for that additional purpose. We must therefore consider carefully the ways in which assessment data are actually used. Paul Newton has identified 22 such uses[5]. These are, however, only broad categories. If one considers each purpose in detail, the number of uses for the data can multiply.

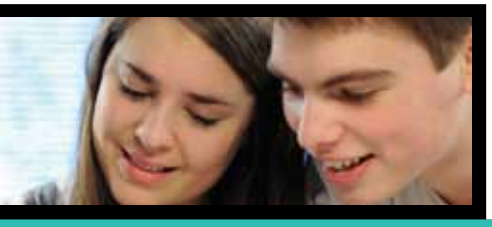
For example, one of these purposes, "Institutional Monitoring", comprises a huge number of uses for assessment data. It is now common in England for data to be used for: school-by-school performance tables; judgments on whether or not schools have hit test and examination targets; performance pay and performance appraisal judgments for teachers and school leaders; assessments of teachers' qualifications for promotion; within-school comparisons of the relative performance of different teachers; judgments as to whether schools qualify for school-by-school schemes including specialist school and training school status; appraisals as to whether they should be threatened with closure; and decisions on whether or not private companies running some schools and local authorities should qualify for incentive payments. Test and examination data are also pivotal in Ofsted inspection judgments[6].

Assessment information has become a proxy measure that is supposed to facilitate judgments on the quality of most elements of our education system: its **teachers**, **head teachers**, **schools**, **support services**, **local authorities** and even the **government itself**. This represents a fundamental change from the situation even 20 years ago, when test and examination results were predominantly meant to serve as indicators of what a **pupil** knew and understood of a subject.

Sometimes, where there is more than one use to which assessment data is being put, it is not clear which is meant to take priority. Thus, for example, in countries where school-by-school performance data are published, this is commonly designed to facilitate judgments both on the quality of the institution and on an individual pupil's progress and future learning needs. But there may be negative consequences for the pupil, if an **institution** takes actions designed to improve its performance in the measured assessments which go against the **young person's** long-term educational needs, for instance, where teachers drill pupils in techniques for earning marks at the expense of teaching for deeper understanding.

5 This list was submitted as part of the evidence from QCA to the inquiry on assessment and testing conducted by the House of Commons Children, Schools and Families Committee in 2007/08. See: http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/169/16906.htm#n35

6 "No school can be judged to be good unless learners are judged to make good progress" [As measured by assessment results]. Ofsted guidance to inspectors, February 2009.



To many policy-makers, it seems attractive, simple and cost-effective to use data from single assessments for multiple purposes – a "no brainer" from a policy perspective. Indeed, David Bell, permanent secretary at the Department for Children, Schools and Families, told MPs in 2008: "While I hear the argument that is often put about multiple purposes of testing and assessment, I do not think that it is problematic to expect tests and assessments to do different things"[7]. However, even where a single set of assessment data is technically fit to be used for multiple purposes – which is certainly not always true – the purposes may still clash in terms of impact and consequences. As argued above, using assessment data for institutional monitoring can have a negative impact upon the quality of education in that institution, which clashes with the most fundamental of uses of assessment data in improving pupils' learning. Clarity about the legitimate and illegitimate uses of assessment data, and the intended or unintended consequences of those uses, is crucial.

For the purposes of further discussion, this commentary simplifies current uses of assessment by clustering them in three broad categories.

1 The use of assessment to help build pupils' understanding, within day-to-day lessons.

2 The use of assessment to provide information on pupils' achievements to those on the outside of the pupil-teacher relationship: to parents (on the basis of in-class judgments by teachers, and test and examination results), and to further and higher education institutions and employers (through test and examination results).

3 The use of assessment data to hold individuals and institutions to account, including through the publication of results which encourage outsiders to make a judgment on the quality of those being held to account.

Another important way to understand the different uses and impacts of assessment is to see the assessment system as a structure which both **provides information** and **influences what people do**. The latter should not be overlooked, as publishing information, which has consequences attached for those who come out well or badly on the basis of these data, will influence teaching, although not necessarily positively.

Governments have sometimes claimed that national tests "drive up standards". Test scores in England have increased significantly since 1995 although the increase has slowed down considerably over recent years. There has been controversy about the extent to which such change reflects only the development of skill by teachers in "teaching to the test". A pertinent example of the issue is provided by research, from Michael Shayer and colleagues, which has shown that pupils' performance on tests of science reasoning has actually declined between 1976 and 2006/7.

The substantive part of this commentary is intended to raise questions about the extent to which assessment systems in the UK are capable of satisfying all the demands that are placed upon them.

Sources:

Newton, P. (2010, in press) Educational Assessment – Concepts and Issues: The Multiple Purposes of Assessment. In E. Baker, B. McGaw, & P. Peterson, (Eds.) *International Encyclopedia of Education. Third Edition.* Oxford: Elsevier.

Shayer, M. & Ginsburg, D. (2009, in press) Thirty years on - a large anti-Flynn effect? (11): 13- and 14-year-olds. Piagetian tests of formal operations, norms 1976-2006/7. *British Journal of Educational Psychology*. Available for free down-load from the British Psychological Society at: http://bpsoc.publisher.ingentaconnect.com/content/bpsoc/bjep/pre-prints/313152;jsessionid=as6c4e0rbt676.alice

Stobart, G. (2008) *Testing Times: the uses and abuses of assessment.* Abingdon: Routledge.

7 David Bell, evidence to the House of Commons Children, Schools and Families committee, 28/01/08

Quality

Introduction

In this section, we map what to look for in an effective system of assessment. In doing so, we divide the discussion into three parts, each of which corresponds to one of our three broad categories of uses of assessment.

One technical note, before we start, is to explain a distinction routinely made by experts in this field: the characterisation of assessment as either **formative** or **summative**.

Formative is the use of day-to-day, often informal, assessments to explore pupils' understanding so that the teacher can best decide how to help them to develop that understanding. Summative is the more formal summing-up of a pupil's progress that can then be used for purposes ranging from providing information to parents to certification as part of a formal examination course.

It should be noted that assessments can often be used for both formative and summative purposes. **"Formative" and "summative" are not labels for different types or forms of assessment but describe how assessments are used.** For example a task or activity is not formative unless the information it provides is actually used to take learning forward. The distinction is undoubtedly useful in helping to understand the different uses of assessment, and so we use it in this section.

Quality in formative assessment/assessment for learning

What a pupil does or says will be observed and interpreted by the teacher, or other learners, who build on that response to develop a dialogue aimed at helping learners to take their next steps. This is formative assessment, which contrasts with summative assessment.

There are characteristic differences between the two uses of assessment:

- Summative comes at the end of learning episodes, whereas formative is built in to the learning process;
- Summative aims to assess knowledge and understanding at a given point in time, whereas formative aims to develop it;
- Summative is static and one-way (usually the teacher or examiner judges the pupil), whereas formative is on-going and dynamic (feedback can be given both to the pupil and the teacher);
- Summative follows a set of pre-defined questions, whereas formative follows the flow of spontaneous dialogue and interaction, where one action builds on (is contingent upon) an earlier one.

The key difference is expressed in the first bullet point above: formative assessment is a central part of pedagogy. This explains why many teachers find it hard to implement; it may challenge them to change what they do, how they think about learning and teaching, and the way in which they relate to their pupils. For example, questions or tasks used in class should be chosen and fashioned in the light of their potential to engage pupils in making contributions that can reveal key strengths and weaknesses in their understanding. What is revealed is often surprising and unexpected.

The most helpful response can often involve steering a discussion in directions that were not envisaged, so that the original lesson plan must be put on hold. Many teachers have said that "I feel I am losing control", but some such "loss" may be essential if the teaching is to respond to the needs of the learners. Research evidence from the TLRP Learning How to Learn project suggests that what teachers need is not rigid lesson plans but frameworks of key ideas that will enable them to maintain the "flow" towards learning goals whilst adapting the lesson to take account of pupils' ongoing struggles or leaps forward in understanding.

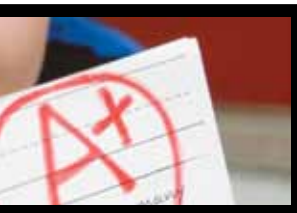
Research by Paul Black and colleagues also shows that summative tests can be used formatively if the pupils' responses are discussed with them to develop an exploration of important aspects that their answers may reveal. The key difference does not lie in the test questions, but in the purpose for which the responses are interpreted and used.

However, frequent summative testing is not, of itself, formative. A teacher may set pupils some questions, whether in a test or in routine written work, and in the light of their results tell them what they need to do to reach, for example, the next target or level. This is not formative unless the interaction is designed to help pupils to learn. This crucial point can be illustrated by a boy's response to the marking of his homework. The teacher had written on the work "use paragraphs", to which he retorted: "If I'd known how to use paragraphs, I would have done". Marks, levels, judgmental comments or the setting of targets, cannot, on their own, be formative. Pupils may need help to know **how** they can improve.

The term "assessment for learning" is often used interchangeably with "formative assessment". In 1999, the Assessment Reform Group defined assessment for learning as "the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there". ARG also identified ten principles for formative assessment practice, arguing that it should:

1 be part of effective planning,
2 focus on how pupils learn
3 be central to classroom practice
4 be a key professional skill
5 be sensitive and constructive
6 foster motivation
7 promote understanding of goals and criteria
8 help learners know how to improve
9 develop the capacity for self-assessment
10 recognise all educational achievement

What, then, is required if the aim of formative feedback is to help learners to grow in capacity to manage their own learning, in order to progress effectively in developing capability in specific subject domains? As implied by the ARG definition, learning can be seen as a journey in which learners need to know where they are at present, where they are going, and how to get from the one place to the other.



This is deceptively simple. It is only in trying to express these things in discussion that most of us come to realise where our difficulties lie: there is wisdom in the familiar saying, “How do I know what I think until I hear myself speak? ”. Likewise, we may need help to see and understand our destination with sufficient clarity that we can guide ourselves in the right direction. Both of these require that we are actively engaged in discussion, of our ideas and of our learning goals. Peer-discussion is essential to achieving such engagement.

TLRP projects investigating how pupils can learn effectively in groups demonstrated that when teachers implement practical group-work strategies, based on the development of social skills, followed by communication skills, followed by problem-solving, significant gains can be made in attainment, motivation and behaviour. It is not just that, through group discussion, all pupils in a class can get involved. There is also benefit in pupils seeing their own ideas reflected in the responses and challenges of their peers, and in learning from them about different ways of tackling the problems that are intriguing. Pupils benefit from doing all of this in the language and style that they use with one another. Yet, as the group-work projects emphasised, the practice of constructive discussion in groups has to be taught. If pupils merely compete to prove who is right, or dismiss one another’s comments rather than taking them seriously, they are again being judgmental rather than formative. Thus the ideal is that pupils engage in formative assessment for one another.

Sources:

Assessment Reform Group (1999) *Assessment for learning: beyond the black box.* Cambridge: University of Cambridge School of Education.

Assessment Reform Group (2002) *Assessment for Learning: 10 Principles.* Cambridge: University of Cambridge School of Education.

Baines, E., Blatchford, P., Kutnick, P., Chowne, A., Ota, C. & Berdondini, L. (2008) *Promoting Effective Group Work in the Primary Classroom. A Handbook for Teachers and Practitioners*. Abingdon: Routledge.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003) *Assessment for learning: putting it into practice.* Maidenhead: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002) *Working inside the black box: assessment for learning in the classroom.* London: King’s College London.

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (under review) Validity in teachers’ summative assessments*, Assessment in Education.*

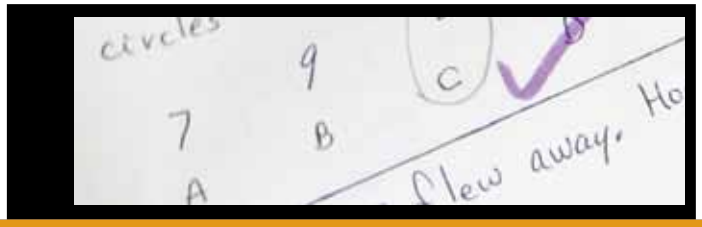
Blatchford, P., Galton, M., Kutnick, P., Baines, E., Berdondini, L., Chowne, A., Hargreaves, L., Ota, C., Page, C., Pell, A., Smith, J. & Steward, S. (2005) *Improving pupil group work in classrooms: A new approach to increasing engagement and learning in everyday classroom settings at Key Stages 1, 2 and 3. TLRP research briefing 11.* London: TLRP.

Christie, D., Tolmie, A., Thurston, A., Howe, H. & Topping, K. (2009) Supporting group work in Scottish primary classrooms: improving the quality of collaborative dialogue. *Cambridge Journal of Education* - Special Issue on Group Work, 39(1): 141-156.

Gardner, J, (Ed.) (2006) *Assessment and Learning*. London: Sage (an Assessment Reform Group publication).

James, M., Black, P., Carmichael, P., Conner, C., Dudley, P., Fox, A., Frost, D., Honour, L., MacBeath, J., McCormick, R., Marshall, B., Pedder, D., Procter, R., Swaffield, S. & Wiliam, D. (2006) *Learning How to Learn: tools for schools*. Abingdon: Routledge.

James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M-J., Fox, A., MacBeath, J., Marshall, B., Pedder, D., Procter, R., Swaffield, S., Swann, J., & Wiliam, D. (2007) *Improving Learning How to Learn - classrooms, schools and networks.* Abingdon: Routledge.



Quality in summative assessment

The characteristics of a summative assessment can vary according to the use made of that assessment, and also according to whether it is marked by the teacher of the pupil who is being assessed, or externally. We therefore divide this section into three:

1. Summative assessment by teachers to be used within their school.
2. Summative assessment by teachers for use externally, outside of their school.
3. Tests and examinations that are marked outside of the school.

Summative assessment by teachers to be used within their school

Teachers routinely sum up what their pupils have learned, because an effective appraisal of the progress pupils have made in their understanding is integral to helping them move on. Yet this is not the only use to which in-class assessment can be put. Its other purposes can include: reporting on a child's progress to parents; helping school managers decide which class a pupil is to be placed in; helping a pupil to decide which subjects to pursue in options choices; forming all or part of an external qualification; and providing information to the outside world on the standards reached by pupils in a particular school.

Reliability and **validity** are central in all types of summative assessment made by teachers.

Reliability is about the extent to which an assessment can be trusted to give consistent information on a pupil's progress; validity is about whether the assessment measures all that it might be felt important to measure.

The case for teacher judgment contributing to the way pupil performance is evaluated has been well rehearsed. In principle, it rests on the argument that teachers can sample the range of a pupil's work more fully than can any assessment instruments devised by an agency external to the school. This enhances both reliability (because it provides more evidence than is available through externally-devised assessment instruments) and validity (it provides a wider range of evidence). Together maximum validity and optimal reliability contribute to the **dependability** of assessments – the confidence that can be placed in them.

It matters that these assessments are dependable. Among the problems that can undermine dependability are unfair or biased marking, variations in the standards applied by different teachers, and failing to reflect important aspects of understanding or skill.

In sketching out ideals for a system of in-class summative assessment, it is, as ever, important to bear in mind the intended and actual (sometimes unintended) uses of assessment information. Teachers will often tailor the characteristics of the assessment to suit its purpose. We develop this argument below.

It is helpful, in this context, to think of in-class assessment in terms of the likely destination for the information it generates. Are the assessment data meant primarily for use within the classroom, beyond the classroom but within the school, or beyond the school?

Quality summative assessment within the classroom calls for:

- pupils to be actively engaged in monitoring their own progress;
- teachers to understand and be able to articulate the nature of the progress being aimed at;
- teachers to be skilled at using a range of methods to assess pupil learning;
- teachers to adopt manageable recording procedures that enable them to keep track of each pupil's learning, without feeling obliged to record everything;
- teachers to be able to communicate effectively with each pupil.

Quality summative assessment across a school calls for:

- manageable expectations of the teacher to report at intervals on the pupils for whom they have a responsibility;
- provision to minimise both the variations in the standards applied by different teachers and the possibility of biased judgments;
- schools to act in a considered way on the summative assessments received from teachers (rather than simply filing them away);
- a sense of audience in the ways in which information about progress is communicated to parents/guardians.

We consider the requirements of quality summative assessment beyond the school in the section below.

In terms of the validity of teacher assessment for use within a school, teachers should think about the quality of the tests and tasks they develop for in-class assessment. Many secondary teachers, for example, use test questions taken from pupils' textbooks, or past examination papers in the years leading up to those assessments. Yet the limitations of the short (thereby economical) tests used in public examinations mean they can threaten the educational values inherent in the subjects tested. TLRP thematic work, led by the Assessment Reform Group, on the Assessment of Significant Learning Outcomes, examined the problem of creating assessments that validly reflect the aims of the curriculum. In many cases the assessment tail continues to wag the curriculum dog.

Aligning assessment with curriculum: the case of school mathematics in England

What should be taught and therefore assessed in school mathematics continues to be controversial. Groups differ in their views of the aims of mathematics education, over the teaching needed to secure these aims, and over the means to assess their achievement. The operational meaning of their aims is often not clear, and the means are often ill thought-out and ill informed. Yet some consensus about what constitutes "school mathematics" is needed if valid assessments are to be developed. In other words there is a need to clarify the "construct" before valid measures can be developed to assess it.

Current views about what school mathematics should be are often quite different. One view is that mathematics is the performance of routine algorithms; another sees mathematics as a tool to tackle "everyday" or "real world" problems. The former leads to assessment of achievement with well-defined exercises, which have a single right answer, with learners inclined to think of achievement as arriving at that answer. The latter looks for evidence of a capacity to tackle the rather messy contexts which are characteristic of every-day problems: problems for which there is no right answer, and where explanation of the way the problem has been defined, and of the approach adopted, is as important as the "answer" itself. Such work is much more demanding to guide, and harder to assess. Yet pupils taught in this way achieve as well in the GCSE as those taught in more traditional methods. They also take more interest in the subject, are better able to see mathematics as useful in everyday life and better able to tackle unusual problems.

The testing system is of course of crucial importance here. With time-limited tests to cover a very full curriculum, any activity that involves much more time than that in which a single examination answer can be given is not possible. Therefore realistic problems are ruled out. This results in an invalidity block, which could in principle be cleared by strengthening the use of teachers' own assessments in national tests and public examinations. The experience of a recent project, which aimed to explore and develop teachers' summative assessments, was that mathematics teachers can develop their own summative assessment in ways that they find rewarding and which can produce dependable results. But such development is difficult to achieve.

Thus, whilst the National Curriculum can be interpreted as reflecting a valid representation of mathematics, the testing system does not realise this potential. To repair this mis-alignment requires changes that demand professional development for teachers, and a consensus about the aims of mathematics education, which does not at present exist.

Sources:

Advisory Committee on Mathematics Education (ACME) (2005) *Assessment 14-19 Mathematics.* London: The Royal Society.

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (under review) Validity in teachers' summative assessments*, Assessment in Education*.

Daugherty, .R., Black, P., Ecclestone, K., James, M. & Newton. P. (2010, in press) Chapter 9: Assessment of Significant Learning Outcomes*.* In R. Berry & R. Adamson (Eds.) *Assessment Reform and Educational Change,* New York: Springer.

Ernest, P. (2000) Why teach mathematics? In S. Bramail & J. White (Eds.) *Why learn maths?* London: Institute of Education.

Routine in-class assessment does not always require the creation of specific tasks or tests. In many cases, it will be appropriate for teachers simply to record their judgments of pupils' progress during day-to-day classroom work and review these at appropriate intervals to come to a summing-up judgment. This applies particularly in primary schools where class teachers have more opportunity than subject teachers in secondary schools to amass a wealth of information about each pupil's achievements.

On reliability, there is less of a need for an assessment to be reliable if it is meant, primarily, to be used within a class, or within a school. There is no need to put in place moderation procedures for routine in-class summative assessments when their main purpose is to provide a platform for further learning. They are only necessary when the data from assessments are to be used externally.

Quality in summative assessment by teachers requires a high level of skill on the part of the teacher. At the same time, teachers need to ensure that their formative assessment practices are not distorted by the demands of summative assessment. This is a challenging agenda.

Summative assessment by teachers for external purposes

Quality summative assessment beyond the school is important, for example, when a pupil moves from one school to another; when the judgments made by a pupil's own teacher contribute towards an external qualification; and when summative assessment by teachers forms part of a wider system of assessment at local, regional or national level.

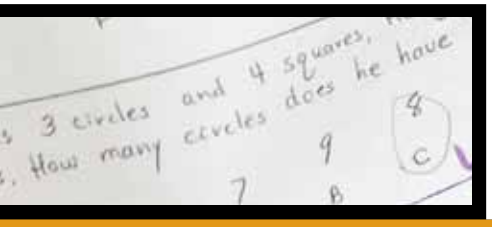
In such contexts, the credibility of the judgments made by teachers will need to be manifestly consistent and unbiased. Systems will be needed to ensure that all teachers engaged in making judgments in a particular context are working in comparable ways to an agreed set of criteria and standards.

For any context in which a much larger number of teachers are making judgments independently of each other, a more sophisticated infrastructure of guidance, training, support and cross-checking is required if the quality of those judgments is to be assured.

For all assessment, pupils, parents and teachers need assurance that the results for a particular pupil are comparable across different teachers in a school and between schools. Few schools are able, by themselves, to audit inter-school comparability, but they should have procedures in place to check intra-school comparability, to be followed up by inter-school moderation. In secondary schools, all subject departments should have a clear and documented assessment policy including specifications for the assessment instruments to be used, for ensuring validity, for resolving differences of opinion between teachers, and for procedures to be used to check the assignment of levels or grades. In primary schools, there should be an agreed assessment policy that provides for teachers to have time for reviewing each pupil's work and discussing the evidence used in assessing it.

The reliability of results becomes more important when assessment data are intended to be used outside of a school to help form judgments on that institution, its teachers, a local authority, a nation or region, or public judgments on the levels of attainment of the pupils. Summative assessment by teachers can and does have a role, around the world, in large-scale systems of assessment. However, there can be a conflict when teacher judgment is used to summarise pupil performance, and when the performance data are then used for accountability purposes. This was highlighted in a review of large-scale assessment systems in seven countries: "...While many systems rely on teacher judgment for assessments that are high stakes for students, there are...no systems that rely on teacher judgment for assessments that are high stakes for teachers"[8]. This needs to be borne in mind if one is seeking to construct systems of public accountability that are based on teacher assessment.

8 Black P & Wiliam D (2007): Large-scale Assessment Systems: Design principles drawn from international comparisons. *Measurement*, 5 (1): 1-53.



Externally marked tests and examinations

What makes for good tests and examinations?

It is difficult to do justice here to the expertise, especially in awarding bodies and test development agencies, which has been built up across the UK, and internationally, in response to this question. Here we simply summarise some of the most important considerations.

Clarity of purpose is essential, in order to judge the validity of the assessment. If there are multiple purposes, we need to consider whether the assessment can meet all of them effectively. The assessment needs to measure what it claims to measure. Examinations that claim to measure analytical reasoning should not be so predictable that they actually become tests of pupils' ability to recall information. A reading test that involves some writing should be designed to reward those who can read well, but who may write badly, rather than those who simply write well.

A test or examination should focus on what matters in the curriculum, rather than simply what is easy to measure. If the test is not measuring what matters in the curriculum, important untested aspects are likely to be downplayed in teaching. One of the reasons given for scrapping key stage 2 tests in science in England from 2010 was that teachers were neglecting crucial hands-on science skills because they were not examined in the written tests.

Cautious interpretation of the results is essential. Those designing and using the tests must be clear what can reasonably be inferred from the results. They should question whether results from a single test can be a reliable and valid key indicator of national standards in that subject, as a whole, particularly if there is extensive teaching to the test. They should consider the likely scale of any measurement error. The results should not be misinterpreted; for example, 20 per cent of pupils in England failing to reach the government's expected level in English does not mean that they cannot read and write, as some media coverage has implied[9].

Most importantly tests and examinations should contribute positively to the pupil's progress in learning, and not undermine good teaching.

Taken together, these are challenging expectations to place on any test, examination or other assessment procedure. Evidence from a small sample of pupil performances on a necessarily limited number of tasks can only aim at a very general judgment of the pupil's attainments in an aspect of learning. Expressed in these terms, imperfections are inevitable. In reality, the best hope of those who design and manage assessments is to minimise imperfections. There is a need for this to be recognised; claims made about what tests and examinations can tell us should be suitably modest.

Some of the policy challenges that arise from these debates about quality are discussed later in this commentary.

Sources:

Assessment Reform Group (2006): *The role of teachers in the assessment of learning.* London: Institute of Education, University of London.

Boud, D. (2002) *The unexamined life is not the life for learning: re-thinking assessment for lifelong learning,* Professorial Lecture. University of Middlesex.

Daugherty, R (2010, in press): Summative Assessment: the role of teachers. In E. Baker, B. McGaw, & P. Peterson, (Eds.) *International Encyclopedia of Education.* Third Edition. Oxford: Elsevier.

Harlen, W. (1994) (Ed.) *Enhancing Quality in Assessment.* London: Paul Chapman Publishing.

Harlen, W. (2004) A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: *Research Evidence in Education.* London: Institute of Education: EPPI centre.

Harlen, W. (2007) *Assessment of Learning*. London: Sage.

Newton, P. (2009) The reliability of results from national curriculum testing in England, *Educational Research*, 51(2): 181-212.

Stobart, G, (2009) Determining validity in national curriculum assessments. *Educational Research*, 51(2): 161-179.

Stobart, G. (2008) *Testing Times: the uses and abuses of assessment.* Abingdon: Routledge

9 "150,000 children unable to read and write at 11" Independent 2/04/09; "A quarter of all children leave primary school unable to read and write", from The Economist: "Let us hope: British schools", 28/01/06.

Quality in accountability

Any decision on the use of assessment data for accountability purposes has to keep two key questions in mind.

1 What are the data attempting to measure?

2 What will be the consequences of publishing the data for accountability purposes?

What are the data attempting to measure?

When assessment results are used for accountability, they can inform judgments on the effectiveness of particular teachers, subject departments, schools, local authorities, the government, other institutions, policies, and on national education systems as a whole.

Several other questions follow. In the case of a school, for example, the key concerns for policy-makers should be whether the use of assessment data for accountability purposes is fair; whether it is valid, in measuring all that one might want to measure; and whether its use supports the inferences that outsiders will make when presented with this information. If the accountability system is meant, in addition, to provide information to influence the actions of teachers and school leaders, a further key question to ask is whether the system identifies opportunities and responsibilities for improvement.

We acknowledge that there are four distinct accountability systems in the four countries of the United Kingdom. Much of the discussion below relates principally to England.

When school-by-school information is presented in performance tables to parents, these data clearly invite a judgment on the effectiveness of a particular institution. Policy-makers will need to ask whether the data support this judgment. Do the aspects of the curriculum that are assessed reflect the entire scope of what is thought to make for an effective institution? Is assessment data for a particular year group reliable enough to yield a dependable judgment as to the institution's quality? And do characteristics outside of the school's control affect the data and its position relative to others with which it is being compared?

Typically, the publication of test and examination data creates a problem: the public routinely draws inferences from assessment results that go well beyond the inferences that the tests are actually designed to support. In terms of performance tables in England, for example, parents want to know which of a few primary schools in their area is likely to educate their child the best. This is a very long way removed from the kinds of inferences that national tests are designed to permit, such as, for example, that Jo has attained more reading proficiency than Sam.

A recent report of an "expert" group of education professionals, who were asked to advise the government in England on the future of key stage 3 assessment, praised the "transparency" created by England's school accountability arrangements. But the system is not transparent about what can legitimately be inferred from the results and what cannot. Furthermore, the report was silent about the existence of error in the measurement of pupils, teachers, schools and national standards.

This latter phenomenon has been well documented. In 2000 it was claimed that at least 30 per cent of pupils in English national curriculum tests might be given the wrong level. This figure has been challenged since, although a study in 2009 suggested that nearly half of pupils might have been given the wrong level in recent national tests for English. In Northern Ireland, research shows that as many as 30 per cent of candidates for the state transfer tests were misclassified.

More broadly, it needs to be acknowledged that a single test result can only ever be a "snapshot" of a child's learning, as the official DirectGov[10] website indicates in a section written for parents in England: "The [national] tests won't give you a complete picture of how your child is doing at school – they provide a 'snapshot', showing how they performed in selected parts of a subject on a particular day".

10 DirectGov website: http://www.direct.gov.uk/en/Parents/Schoolslearninganddevelopment/ExamsTestsAndTheCurriculum/DG_4016665.

Given the weight now being placed on school-by-school information, and the inferences being made on the basis of test and examination data, policy-makers have a responsibility to the public to be clear about what the results can tell them with confidence, and what they may suggest with less certainty. Those compiling and presenting information on each institution should provide a disclaimer, stating what the data are designed to measure, and, most importantly, what they cannot. Conventional school examination data, then, should be published with this "health warning": that they measure only part of a school's business and that these results are influenced by factors beyond the school's control, such as the prior attainment and background characteristics of its pupil population. Consideration should be given, also, to acknowledging the existence of error in the marking of tests and examinations. This principle could be extended to other aspects of the school accountability system, such as the use of pupil results to judge teacher performance. Overall, the publication of data, which carries major consequences for adults as well as children, should be seen as calling for responsibility, and even humility, from those designing the measures and publishing the data. All such information is provisional; it rests on assumptions and it can never be definitive.

Especial care should be exercised when using unadjusted or "raw" pupil data to make judgments about the quality of particular local authorities, schools, or teachers. The recent National Challenge scheme in England, in which schools were publicly challenged to improve the percentage of their pupils achieving five or more GCSE A* to C grades including English and mathematics, carried particular risks. In launching the scheme the government stated that institutions that had below 30 per cent of their pupils achieving this benchmark were underperforming. However, the data used do not support this inference.

Pupils' GCSE performance is affected by some factors that the school can control, such as the quality of teaching; and some, which usually it cannot, such as pupils' prior attainment levels and their home backgrounds. A school with results below this benchmark might be educating better, given the factors it can control, than one with results above it. More sophisticated measures, which seek to isolate the responsibility of the school for its pupils' examination performances, are needed to support such inferences.

In England, some performance tables and analyses designed for self-evaluation do, however, contain contextual value-added scores that take some account of pupils' prior attainment and social background. In Scotland many local authorities have provided individual schools with comparative information about the nationally defined Education 5-14 levels of attainment, which cover primary education and the first two years of secondary. In self-evaluation, which may or may not include interaction with local authority staff, each school can compare the proportion of its pupils attaining each level with the local authority and the national average. (In recent years there has been no collection of national data, so the national averages provided by local authorities have typically been slightly out of date.) Some large local authorities, which have enough schools to make this possible, also enable a school to compare itself with the averages for a group of schools with similar characteristics. Each secondary school has on-line access (on the Scottish Government Statistics website) to benchmarking information that compares its post-16 attainment levels with national averages and with the performance of other schools with similar characteristics. Schools are encouraged to use this information in self-evaluation. In an inspection, HMIE uses it as the basis of discussion with the school about its own performance.

In both England and Scotland this "contextualized" approach may appear to be more just – and perhaps more useful to a school's self-evaluation - in that schools in similar circumstances are compared. But inbuilt assumptions still need to be questioned. For example, do similarities across a limited number of indicators suggest similar schools? Does comparison really lead to improvement? Given the evidence on the importance of within-school differences, rather than cross-school differences[11], are detailed school-level analyses really so useful?

Policy-makers sometimes claim that the publication of national test results for each school is a value-neutral act of sharing information with the public. If the public then, on the basis of these data, makes judgments on the overall quality of each school, policy-makers might argue that this was not their intention. But this is disingenuous. Information will always be used in one way or another. A better system would acknowledge that there is a need to ensure that the public understands what these data can tell them, and what they cannot.

11 Such differences are illustrated in the findings of the TLRP Learning How to Learn project.

What will be the consequences of publishing data for accountability purposes?

To ask this question is to acknowledge that the act of revealing information about a particular institution will have implications, both intended and unintended.

Any high-stakes indicator will exert pressure on those involved and thereby influence their work. Accountability pressures have led schools to become more focused on some aspects of what policy-makers may see as important in education, such as the securing of good grades in key examination subjects for their pupils. However, in doing so, they have also encouraged schools to narrow the range of their aims, to teach to the test, or to choose subjects for which qualifications appear easier to secure. In many such ways, the accountability system may come to damage the very outcomes that it was designed to improve.

Policy-makers should also acknowledge that the publication of data will influence behaviour within institutions, in intended and unintended ways. They should carry out investigations into these effects, ideally in trials before nationwide publication of the information. If unintended consequences are present, this should be acknowledged and the question asked whether their existence outweighs the benefits of collecting and presenting the data. If a decision is taken to press on with publication, policy-makers should ask how its worst effects can be mitigated and take action to minimise negative effects.

Finally, policy-makers should be sufficiently open-minded to consider the accountability system in its entirety and ask whether it is effectively fulfilling the goals that have been set for it. In doing so, they need to ask fundamental questions, which are beyond the scope of this commentary, about the effect their regime is having on trust in the teaching profession. There are alternative, more intelligent, forms of accountability, perhaps as advocated by Baroness Onora O'Neill, a professor of philosophy at Cambridge University and President of the British Academy. In the third in her 2002 series of Reith Lectures she argued that:

"The pursuit of ever more perfect accountability provides citizens and consumers, patients and parents with more information, more comparisons, more complaints systems; but it also builds a culture of suspicion, low morale and may ultimately lead to professional cynicism, and then we would have grounds for public mistrust. In contrast, intelligent accountability concentrates on good governance and an obligation to tell the truth. I think [Parliament] has to fantasise much less about Herculean micro-management by means of performance indicators or total transparency. If we want a culture of public service, professionals and public servants must in the end be free to serve the public rather than their paymasters."

Advocating such a change does not mean calling for less accountability, but rather that it should be viewed and managed differently, with an emphasis on independent governance of institutions by properly qualified people who command public, especially local, support, rather than micro-managed, at too great a distance, by politicians and civil servants.

It also needs to be remembered that the focus on accountability, if that implies holding schools to account mainly for test and examination performance, may detract from the central purpose of any education system – to improve sustained learning and the rounded education of young people. There is strong evidence, referred to in the above section on quality in formative assessment, to suggest the kind of processes that will improve learning. National policy communities willing to build such processes into the design of their improvement programmes are more likely to achieve their aspirations of real educational improvement.

Sources:

Department for Children, Schools and Families (2009) *Report of the expert group on assessment.* London: DCSF.

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

Mansell, W. (2007) *Education by Numbers: The tyranny of testing.* Politico's.

O'Neill, O. (2002) Called to Account, Lecture Three, *BBC Reith Lectures, A Question of Trust.* Accessed at http://www.bbc.co.uk/radio4/reith2002

Qualifications and Curriculum Authority (2009) *Research into marking quality: studies to inform future work on national curriculum assessment.* London: QCA.

Wiliam, D. (2000) Validity, reliability and all that jazz, *Education 3-13*, 29(3): 9-13.

Four pressing challenges for policy-makers

Putting effective in-class assessment into practice system-wide

How effectively is formative assessment, sometimes known as assessment for learning, working in UK classrooms?

This question has preoccupied researchers for many years. The largest study of its kind, the TLRP Learning How to Learn in Classrooms, Schools and Networks project, researched how such practices were developed by teachers in 40 English primary and secondary schools. All were given some training. Then it was investigated how they implemented assessment for learning in practice, and what characterised the schools where this was most effective .

The research lasted from 2001 to 2005 and included surveys of 1,200 school staff and 4,000 pupils, and detailed interviews and observations of 37 teachers' lessons (27 were filmed). It found that what was defined as the "spirit" of assessment for learning was hard to achieve. Although many teachers used techniques associated with assessment for learning, such as sharing success criteria or increasing "thinking time", few did so in ways that enabled pupils to become more autonomous learners. This is a defining characteristic of assessment for learning and learning how to learn. Some 20 per cent of teachers were, however, identified as capturing its spirit, which showed that it is possible.

The "spirit" and the "letter" of assessment for learning

The TLRP Learning How to Learn project drew a distinction between lessons that adopted surface procedures and those that truly captured the deeper principles of formative assessment.

Two English teachers were filmed teaching separate Year 8 classes (13 year olds). They were both attempting to do similar things in similar contexts. In both lessons, the teachers shared the criteria with the pupils by giving them a model of what was needed. The pupils then used those criteria to assess the work of their peers. These elements are central to assessment for learning.

In lesson A, pupils looked at a letter they had written based on a Victorian short story. The teacher modelled the criteria by giving the pupils a piece of writing that was full of errors. They were asked to correct it on their own. The teacher then went through the corrections with the whole class before asking them to read through and correct the work of their peers.

In lesson B, the pupils were asked to consider a dramatic rendition of a nineteenth century poem. The teacher and the classroom assistant performed the poem to the class and invited the pupils to critique their performance. From this activity the class as a whole, guided by the teacher, established the criteria. These criteria then governed both the pupils' thinking about what was needed when they acted out the poem themselves and the peer assessment of those performances.

Lesson A was an example of assessment for learning being followed to the letter, because pupils were only being taught how to guess what the teacher was looking for in correct answers. Lesson B followed the spirit of assessment for learning, because the sequence of activities helped them to learn how to learn for themselves.

Source:

Marshall, B. & Drummond, M-J. (2006) How teachers engage with assessment for learning: lessons from the classroom, *Research Papers in Education*, 21(2): 133-149.

The difficulties many teachers in England have conducting effective in-class assessment have been extensively documented by Ofsted. In 2008, it reported: "Too many teachers still fail to see the connection between accurate and regular assessment and good teaching which leads to learning"[12].

The government in England is attempting to tackle this problem by introducing a £150 million three-year strategy to promote what it bills as Assessment for Learning. Will it work? Analysis of some of the challenges which the introduction of authentic assessment for learning has faced suggests a cautious response to that question.

Evidence from the TLRP Learning How to Learn Project has shown that many teachers are committed to the ideals behind assessment for learning. For example, of those surveyed, 95 per cent said they believed it was important or crucial that their pupils were helped to think about how they learn best; 93 per cent said so in connection to them helping pupils to assess their own learning; and 83 per cent said so in relation to pupils being helped to plan the next steps in their learning.

However, in general many fewer teachers reported actually adopting these techniques in practice. Only 63 per cent reported often or mostly helping pupils to think about how they learn best; only 69 per cent said so in connection with helping pupils to assess their own learning; and only 46 per cent said so in relation to helping them plan their next steps. "It is clear...that teachers' classroom assessment practices contravene their values," said the study.

The Learning How to Learn project, along with other research, suggests at least five reasons.

First, teachers' ideals of helping pupils to become more autonomous learners, who are able to reflect on the development of their understanding, often run up against the reality of the need to raise their pupils' **performance** on the next summative test. The purpose of using assessment in-class to improve understanding often conflicts with its use as a monitoring device by others. The Learning How to Learn project identified three distinct factors in classroom assessment practice: "making learning explicit"; "promoting learning autonomy"; and "performance orientation". The first two were seen as intrinsic to assessment for learning; the third was not. While all were seen to be important by teachers, the research found that they tended to prioritise "performance" more than they believed was appropriate, while they gave the first two factors less emphasis than they would have liked.

Some of the difficulty arises from a lack of understanding of the difference between "performance" and "learning". The eminent psychologist, Carol Dweck, distinguishes two co-existing orientations. A performance orientation is about "looking smart" and avoiding looking dumb; a learning orientation is about "getting smarter" by learning new skills, mastering new tasks, or understanding new things. An over emphasis on performance goals can drive out learning. In England, this has implications for the over-emphasis on ticking off levels and sub-levels in national curriculum attainment targets. Levels that are supposed to be indicators of learning become the goals themselves and can be "clocked up" without any guarantee that the underlying learning has occurred.

Second, there are pressures on teachers' **time**, including the need to cover all aspects of the national curriculum. This means there may be limited opportunities for pupils to develop the self-reflective capacities that are central to assessment for learning. As one teacher told the Learning How to Learn project researchers: "In the main, we're trying to put too much into the curriculum and not allowing the children to explore and enjoy fewer things and make stronger links...we're galloping through".

12 Ofsted (2008) Memorandum to House of Commons Children, Schools and Families inquiry into Assessment and Testing, 2007/08.

Third, there is a evidence of a **"tick-box culture"**, in which assessment information can be seen as being mainly concerned with meeting a bureaucratic need to provide evidence of learning to school managers and others. A primary teacher told the Learning How to Learn researchers: "We are expected, in each subject, to have literally a list of the children and ticks next to them whether they can do this, this and this....The time that you spend filling these in surely can be put to better use in identifying children with problems and then, you know, meeting these needs where they are". The recent Assessing Pupils' Progress initiative, in England, risks encouraging this constant monitoring of pupils' levels and sub-levels although it claims to be promoting assessment for learning as part of pedagogy.

The need to develop effective formative assessment/assessment for learning without it becoming overly bureaucratic is therefore a key, but not insurmountable, challenge.

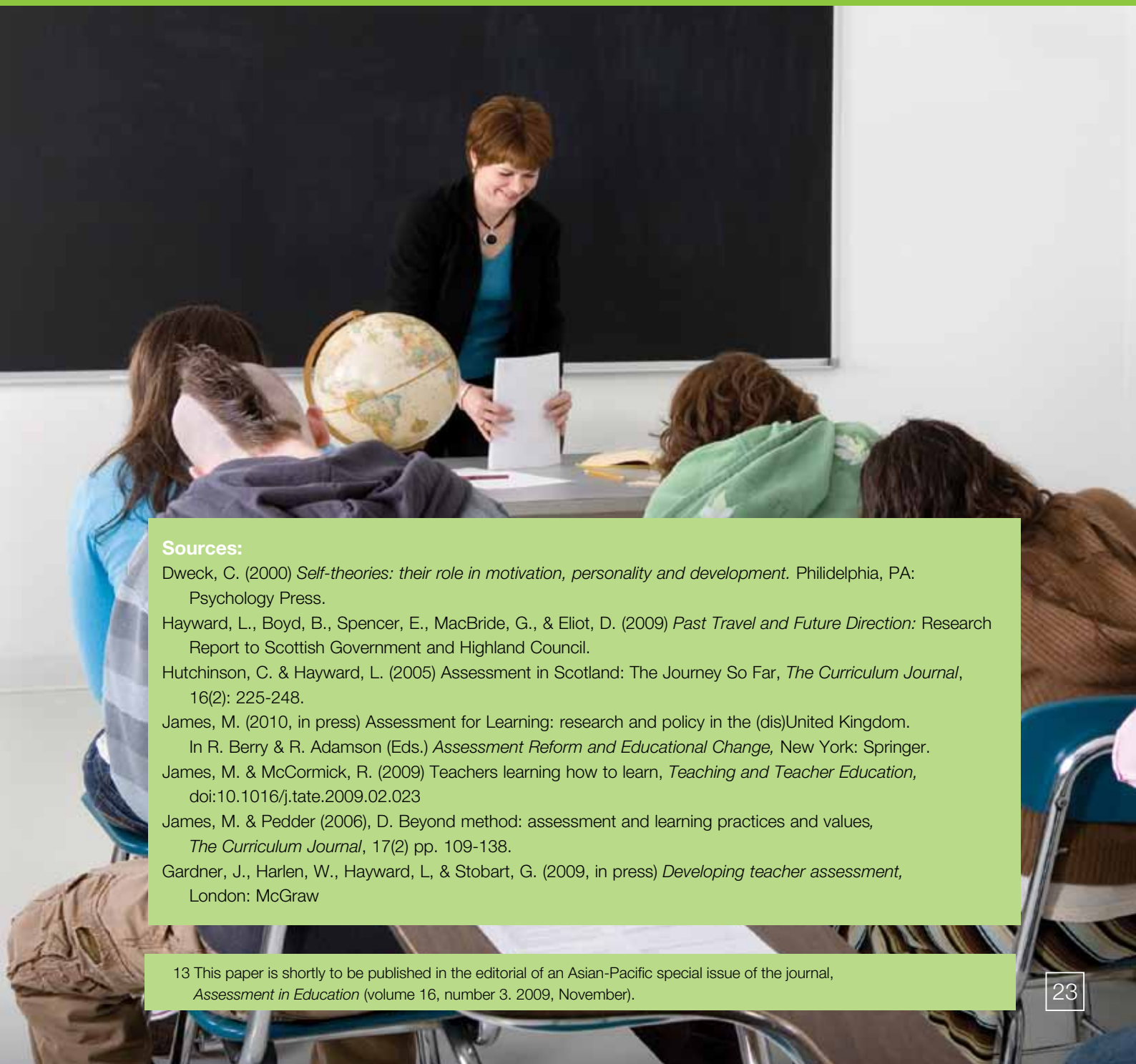
A fourth problem has been called "**scaleability**", or the difficulties policy-makers face in expanding small-scale initiatives, implemented in a few schools with intensive support, into larger reforms. Assessment for learning can be demanding, and while evidence suggests it has proved powerful in research projects with committed teachers supported by academics, further expansion is not simple. This was the problem that the Learning How to Learn project specifically sought to address. A key finding from the project was that **those teachers who were most successful in promoting assessment for learning were also those who engaged in professional development based on collaborative, classroom-focused inquiry actively supported by school leaders**. Another was that teachers develop and communicate new knowledge about effective practice through networking within their school and with other schools. Even short exchanges and occasional meetings can have a positive impact; support for networking does not have to be expensive or hugely time-consuming. However, it does need to be planned and supported so that teachers, as well as pupils, have opportunities for quality learning.

A fifth problem is the lack of attention given to teachers' assessment practices in **initial teacher education** and other training. The support of the government in England for a new professional status, a chartered educational assessor who has undergone extensive training, may go some way to addressing this difficulty but it needs to encompass both the formative and summative purposes of assessment practice.

The honest answer to the question of how formative assessment is working, at least in England, is that effective practice is still patchy. Teachers who generally appear to have a strong idealistic commitment to the thinking behind these concepts often struggle to put them into practice in the face of competing pressures on their time and priorities. This contrasts with the situation claimed of other countries of the UK. These have reduced some of the critical pressures by rejecting whole cohort testing, as the basis of accountability, and promoted assessment for learning through rather different kinds of development programmes. For example, the extension of a Thinking Skills and Assessment for Learning development programme in Wales is based on: close partnership working between civil servants, local authorities and schools; local and national networking to encourage adaptation and spread good practice; and funded reflection and planning time for teachers. A specific intention is to use the professional networks already established with international researchers and collaborators to enhance the programme.

In Scotland, the Assessment is for Learning (AifL) programme has sought to develop a coherent assessment system, paying attention to assessment for formative and for summative purposes. Policy-makers, researchers and practitioners have worked together in networks to develop assessment policy and practice. As a result, AifL is highly regarded in Scotland and evaluations of the programme have stressed high levels of commitment and engagement amongst teachers and learners. In 2003 the then Education Minister described it as a quiet revolution in Scottish education. Recent studies in Scotland suggest that AifL is making a positive impact on teaching, learning and attainment, including in Higher Examination results.

In 2009, assessment experts from Canada, the United States, the United Kingdom, continental Europe, Australia and New Zealand, met at a conference to examine the opportunities and pressures facing teachers who are trying to implement assessment for learning in their classrooms. Confusion over how the term is interpreted appears to be widespread, internationally, and exacerbated by some policy-makers appropriating and using it in ways that contradict its true intentions. In particular, "deciding where the learners are in their learning, where they need to go, and how best to get there" has been taken to mean frequent testing of levels achieved and then focusing on the deficiencies in order to meet the next level. That is, the levels become the goals rather than the learning. The conference participants therefore wrote a position paper to help to clarify the central ideas and to re-emphasise the purpose as enhancing ongoing learning.[13]



Sources:

Dweck, C. (2000) *Self-theories: their role in motivation, personality and development.* Philidelphia, PA: Psychology Press.

Hayward, L., Boyd, B., Spencer, E., MacBride, G., & Eliot, D. (2009) *Past Travel and Future Direction:* Research Report to Scottish Government and Highland Council.

Hutchinson, C. & Hayward, L. (2005) Assessment in Scotland: The Journey So Far, *The Curriculum Journal*, 16(2): 225-248.

James, M. (2010, in press) Assessment for Learning: research and policy in the (dis)United Kingdom. In R. Berry & R. Adamson (Eds.) *Assessment Reform and Educational Change,* New York: Springer.

James, M. & McCormick, R. (2009) Teachers learning how to learn, *Teaching and Teacher Education,* doi:10.1016/j.tate.2009.02.023

James, M. & Pedder (2006), D. Beyond method: assessment and learning practices and values, *The Curriculum Journal*, 17(2) pp. 109-138.

Gardner, J., Harlen, W., Hayward, L, & Stobart, G. (2009, in press) *Developing teacher assessment,* London: McGraw

13 This paper is shortly to be published in the editorial of an Asian-Pacific special issue of the journal, *Assessment in Education* (volume 16, number 3. 2009, November).

Enhancing confidence in tests and examinations

As the section on purposes set out, assessment data, for the most part based on pupil performance in tests and examinations, are now used in an extraordinary variety of ways, underpinning not just judgments of pupils' progress, but helping measure the performance of their teachers, schools and of the nation's education system as a whole, among other uses. These uses can have far-reaching consequences for those being judged by the data.

A first important question for research, policy and practice, therefore, should be the **reliability** of this information. In simple terms, how accurate is the assessment data generated, particularly, through national tests and examinations, as an indicator of what we might want to measure?

Research suggests that we should treat national test results in England, as measures of pupil performance, with caution. As noted above, Dylan Wiliam estimated in 2000 that at least 30 per cent of pupils could be misclassified in these tests. In December 2008, Paul Newton suggested a figure closer to 16 per cent. In March 2009, the Qualifications and Curriculum Authority published research based on tests taken in 2006 and 2007. This analysed the number of times markers agreed on the level to award to pupils' answers in the now-discontinued key stage 3 tests in English, maths and science. The extent of agreement varied from 95 per cent in maths to 56 per cent in English writing, the latter suggesting that markers disagreed on the "correct" level to award in nearly half of these cases.

Meanwhile the public examinations regulator in England, Ofqual, is conducting a programme of research that will investigate the reliability of GCSEs, A-levels and other qualifications.

All of the above may underline the need for published health warnings around the reliability of the tests and examinations, as argued for in the section on quality. The question of the impact on public confidence, of being open about the degree of error in the testing system, needs of course to be taken seriously. However, the argument that there is a need to be transparent about the limits and tentativeness of the judgments being made about individuals under the testing regime carries greater weight. There is also a clear need for more research on reliability.

The second key question surrounds the **validity** of national tests and examinations: do they measure the aspects of education which society feels it is important to measure?

There has been a continuing debate around the validity of national curriculum tests, particularly in England. For example, the absence of any assessment of oracy within the external key stage 1, 2 and 3 English tests has been a source of contention for many, especially as speaking and listening is often seen as the foundation for children's mastery of the subject. If English national test data are central to the construction of performance information on schools but leave out this central part of the subject, how valid are the judgments that follow? Similar arguments by science organisations in relation to key stage 2 science tests – that they failed to measure all that was important about that subject, including experimental and investigative work – are said to have helped persuade the government in England to scrap these assessments.

The third key question surrounds the impact of publishing information on pupils' test and examination scores on classroom practice. This is, arguably, the defining question of the government in England's education policies, and an extraordinary amount has been written about it. There is little doubt that policies such as performance tables, targets and Ofsted inspections that place great weight on test and examination data have driven behaviour in schools. Advocates of this system argue that it has helped teachers focus on what those designing the performance measures regard as essential to the educational experience, such as facility with English and mathematics.

Yet there is now a great volume of material cataloguing the educational side-effects of a structure which is too focused on performance indicators. These include the often excessive and inequitable focus of many schools on pupils whose results may be key to a school hitting particular achievement targets; the repetition involved in months of focusing on what is tested and on test practice, which also serves to narrow the curriculum; and the consequent undermining of professional autonomy and morale among teachers.

The impact on pupil motivation to learn is an area of particular interest. If one of the central aims of assessment for learning is to encourage independent motivation to understand among pupils, findings from research that learners in high-stakes testing systems can become dependent on their teacher to guide them towards answers should be taken seriously.

More generally on pupil motivation, the most extensive review of research in recent years[14] on the effect of the tests found that those that were seen as "high stakes" de-motivated many children. Only the highest attainers thrived on them, with many showing high levels of anxiety. After the introduction of national testing in England, the research found, self-esteem of young children became tied to achievement, whereas before there was little correlation between the two. This effect was found to increase the gap between low- and high-achieving pupils, with repeated test practice tending to reinforce the poor self-image of the less academic. The review also found that pupils tended to react by viewing learning as a means to an end – the pursuing of high marks – rather than as an end in itself.

On the other hand, it has been argued that some children and young people thrive in the face of stressful challenges and that external tests and examinations do motivate pupils to take what they are learning seriously. Indeed, the government in England has suggested this recently[15], when arguing that the possible introduction of new tests, which key stage 2 pupils could take up to twice a year, could increase motivation. There is as yet little evidence for this claim, and it needs further research.

Sources:

Assessment Reform Group (2002) *Testing, Motivation and Learning.* Cambridge: University of Cambridge Faculty of Education.

Black, P., Gardner, J. & Wiliam, D. (2008) *Joint memorandum on reliability of assessments.* Submitted to the House of Commons, Children, Schools and Families Committee: Testing and Assessment. Third Report of Session 2007-2008. Volume II. HC169-II. Norwich: The Stationery Office. Ev 202-5.ISBN 978 0 215 52041 8 (Discussed in Vol. I pp22-6).

Harlen W., & Deakin Crick R. (2002) A systematic review of the impact of summative assessment and tests on students' motivation for learning. In: *Research Evidence in Education Library.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Newton, P. (2008) *Presentation to the Cambridge Assessment Forum for New Developments in Educational Assessment.* Downing College, Cambridge. 10 December.

Qualifications and Curriculum Authority (2009) *Research into marking quality: studies to inform future work on national curriculum assessment.* London: QCA.

14 Harlen W., & Deakin Crick R. (2002) A sytematic review of the impact of summative assessment and tests on students' motivation for learning. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London

15 Department for Education and Skills, (2007) *Making Good Progress: How can we help every pupil to make good progress at school?* London: DCSF.

Justifying the costs of assessment

Extraordinary sums of money are now devoted to our assessment systems in the UK. The key question is whether these resources could be better spent.

In 2005, the consultants PricewaterhouseCoopers published a report, based on an investigation carried out in late 2003, which estimated the annual total cost of the English examinations system as £610 million. This total consisted of £370 million which was spent by schools, colleges, awarding bodies and the Qualifications and Curriculum Authority on direct examination costs, and a further £240 million estimated as the cost in terms of staff time running examination activity in schools and colleges.

More recent information suggests this figure may underestimate today's overheads. A report, in 2009, for the examinations regulators for England, Wales and Northern Ireland collated the incomes of 12 leading awarding bodies covering these countries for the three years to 2007. These figures gave an average yearly expenditure of £659.3 million, an increase of 15 per cent over the previous two years.

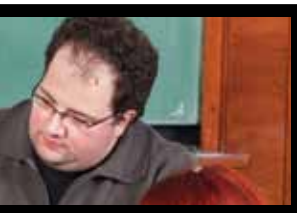
This, however, relates to only a subset of assessment costs: external examinations for the over-14s. Another major type of external assessment, in England, is national curriculum testing. Information from the Qualifications and Curriculum Authority estimated that the total cost of key stage 1, 2 and 3 tests in 2008 was £50.6 million[16].

Trying to get an estimate of the cost of less formal in-class assessment is much more difficult. However, in 2008 the Department for Children, Schools and Families introduced its Assessment for Learning Strategy, which it said would costing £150 million over three years, or £50 million per year. Adding up awarding body income, the direct costs of national curriculum tests and the government's Assessment for Learning strategy in England, then, one arrives at a figure of more than £750 million per year.

To decide whether it is money well spent, one needs to analyse what the assessment system is seeking to achieve, and how it is trying to do it. The full detail required in such an analysis is beyond the scope of this commentary. However, two points are worth making.

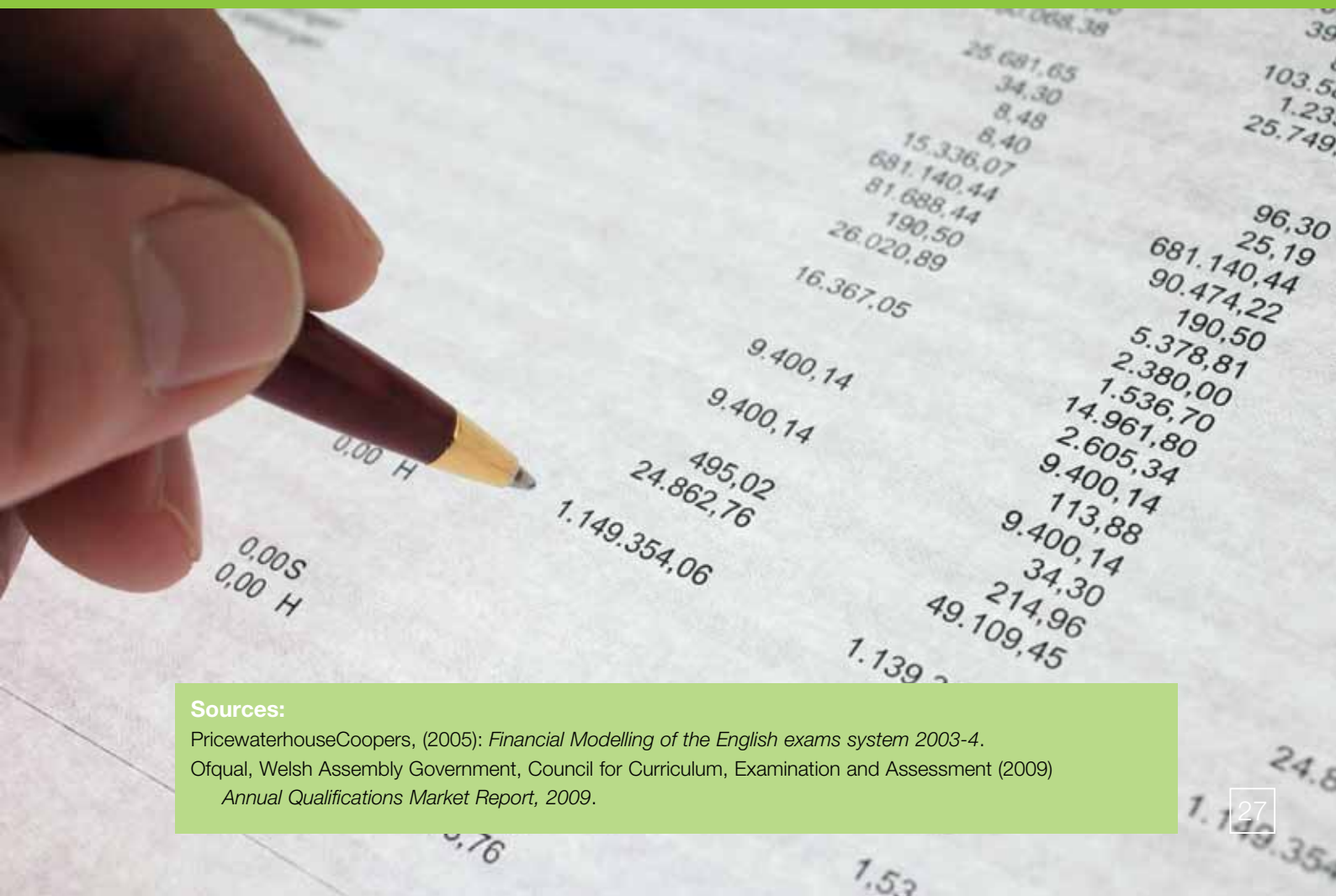
The first is that just because the sums involved appear large, it does not follow that external assessment is necessarily more expensive than alternative forms of assessment and accountability. For example, national assessments using paper and pen tests may well be cheaper than face-to-face teacher-moderated assessments of, for example, pupils' ability to carry out scientific experiments, play musical instruments or make oral presentations, (although the validity of inauthentic assessments should be questioned).

16 Included in a Department for Children, Schools and Families response to a Freedom of Information request, September 2008.



Furthermore, using test results as the focal points of Ofsted inspectors' checks on school quality may be thought to lessen the need for more expensive visits to schools to observe teachers' lessons. In both cases, the current external test and examination regime could be seen to be cheaper than possible alternatives. However, summative assessment performed by teachers may have additional benefits that make it a particularly good investment. For example, teachers discussing with colleagues the grading criteria for such marking, and whether these criteria have been appropriately applied, can be a powerful learning process for professionals. As the TLRP Learning How to Learn project showed, teachers' own learning, with colleagues, was the single most important factor associated with their capacity to promote independent learning in their pupils through effective in-class assessment practices.

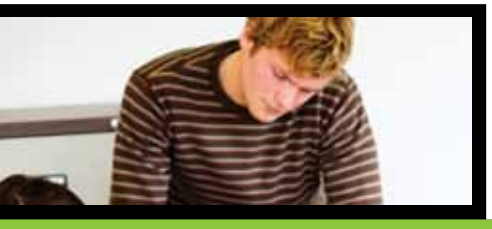
The second point is simpler. There is undoubtedly a great deal of external assessment in English schools. Notwithstanding the government's recent decisions to scrap the key stage 3 tests and the key stage 2 tests in science, the final four years of pupils who stay on to the sixth form are dominated by examinations. GCSEs, AS levels and final A-level papers mean that few terms are now left free of revision and formal assessment. The rise of modular GCSEs is poised to accentuate this trend. Ultimately, this examining load and the financial costs associated with it should be evaluated in terms of educational value in relation to money spent.



Sources:

PricewaterhouseCoopers, (2005): *Financial Modelling of the English exams system 2003-4*.

Ofqual, Welsh Assembly Government, Council for Curriculum, Examination and Assessment (2009) *Annual Qualifications Market Report, 2009*.



Avoiding micro-management

Perhaps few people are aware of the extent to which governments can control the detail of how pupils are assessed. In some countries they exercise that control more than in others.

To sceptics, the story of recent assessment policy in England has been one of ministers and civil servants intervening to make decisions that may later encounter problems of implementation. Some of these problems might have been predicted by better attention to research and previous experience.

In the case of public examinations, there is no set formula for government intervention. In England, the Qualifications and Curriculum Authority – now Ofqual and the QCDA – has official control of the detail of how pupils are assessed. For instance, Ofqual publishes guidelines – codes of practice and criteria – for awarding bodies to follow in developing and administering qualifications including GCSEs and A-levels. However, in practice, ministers have exercised extensive powers, partly through letters of remit sent to the QCA requiring it to investigate and develop particular assessment changes. In recent years, the Department for Children, Schools and Families has: pushed for the introduction of new "functional skills" qualifications in English, maths and ICT; sought to introduce optional harder questions into A-levels in support of a new A* grade; and specified to the QCA the rules for grading the new diploma qualification.

In 2008, the head of a major awarding body complained that the government had ordered the use of calculators in and out of examinations seven times in the past decade and pointed out that rules for the calculation of the A-level A* grade boundary put forward by the QCA were subject to the approval of the Secretary of State[17].

In national curriculum testing, the government, while not setting individual questions or the number of marks required for pupils to achieve particular levels, controls other aspects including which subjects and which pupils are to be tested[18].

As is implicit in earlier sections of this commentary, the government in England, in contrast to Wales, Scotland and Northern Ireland, has developed a version of assessment for learning which shares little of the "spirit" of the definition and principles from the Assessment Reform Group, although the documentation quotes them. Indeed, Assessing Pupils' Progress, the in-class assessment system that is a part of the government's version of assessment for learning in England, is more to do with specifying frequent summative assessment than formative assessment.

Government direction and involvement is not, of course, a new phenomenon. While no-one would contest the right of elected politicians to determine overall assessment policy, their involvement in specifying technical details of assessment models and procedures raises questions over whether they, and some of their advisers, are sufficiently qualified to do so at such a detailed level. Two recent examples underscore this difficulty.

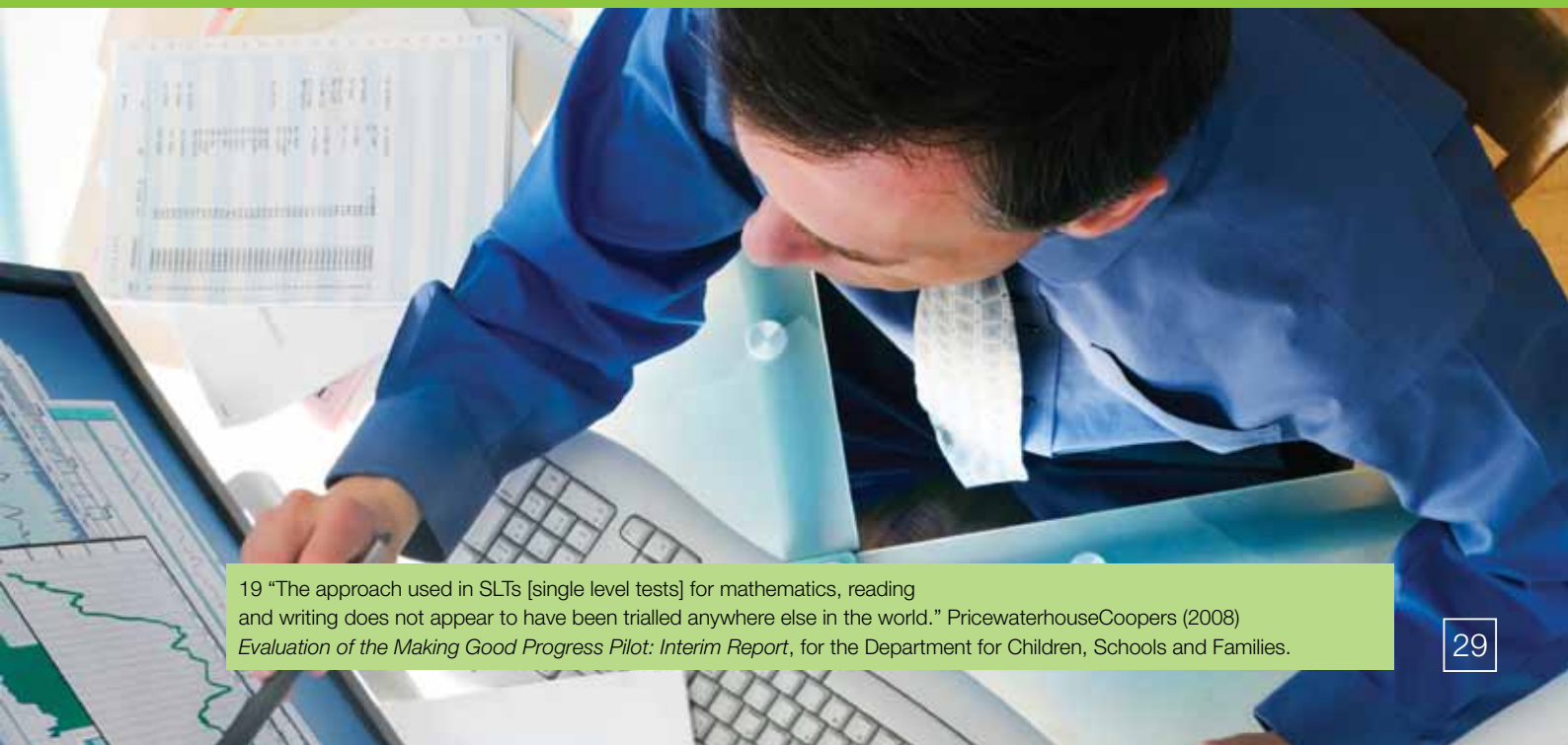
17 "Cambridge exam chief warns interference by ministers is undermining qualifications", Guardian 22/07/08

18 Ken Boston, former chief executive of the Qualifications and Curriculum Authority, told MPs in April 2009 that the Government controlled all aspects of the national curriculum testing process other than the detail of test questions, the marking of the tests and the setting of mark boundaries. See: http://www.publications.parliament.uk/pa/cm200809/cmselect/cmchilsch/c205-ii/c20502.htm

First, single level tests (SLTs), which may eventually replace conventional national curriculum testing in England, is a potentially significant initiative that was specified, in advance of development and piloting, by the Department for Education and Skills. Rather than simply setting a goal for policy, such as that assessment should become more personalised to the needs of the learner, and then letting assessment experts put forward and trial methods to achieve this, the basic characteristics of the testing model were set out from the beginning. Most notably, the Department required: that the SLTs should be substantially shorter than the existing (multi level) tests; should assess only a single level; and should be administered during the programme of study, rather than once the programme of study had been completed. Tests of this sort would be unique to England[19] and are of questionable legitimacy, from a technical perspective. The original intention that this model should span both primary and secondary education has been abandoned (SLTs will not be used with secondary pupils) for reasons that are at least partly technical.

Second, in the case of functional skills tests, government officials specified these should be based upon a "mastery" model of assessment, and be integral to GCSEs in English, mathematics and IT. This was supposed to ensure that pupils who achieved grade C or higher in these subjects had mastered all of the component skills that were deemed necessary to function in the modern world. England has tried to apply the concept of mastery assessment on numerous occasions – in the build-up to GCSE, in GNVQ, in early national curriculum tests – and has failed on each occasion. Mastery assessment can work, when aligned to a mastery model of teaching and learning, but this is not what happens in most schools in England. Without a radical change in approach to teaching English, mathematics and IT, it was clear – and clear right from the outset – that functional skills tests, based on a mastery model of assessment, would result in a substantial drop in the percentages of pupils awarded grade C or above in these subjects. Four years after the tests were first proposed, the government has recently dropped plans to make passing them a requirement for a GCSE grade C.

To be fair, there are also examples of ministers taking seriously the advice of the experts, with evidence from the QCA being used as the basis for policy in spite of initial government wishes to the contrary. Politicians rightly want to hold the education service to account, and to guide overall policy. The assessment community undoubtedly has detailed expertise. Yet, in England, there clearly is a need for a better-defined relationship between these two groups of people. A debate on where political guidance is appropriate, and where it is not, would be a useful step forward in enhancing the trust of the public and education professionals.



19 "The approach used in SLTs [single level tests] for mathematics, reading and writing does not appear to have been trialled anywhere else in the world." PricewaterhouseCoopers (2008) *Evaluation of the Making Good Progress Pilot: Interim Report*, for the Department for Children, Schools and Families.

Conclusion: enhancing public understanding for wise decision-making

UK assessment systems have become increasingly complex. No longer are they simply designed to accredit the achievements of school pupils and contribute to their education; they are expected to serve an enormous range of other purposes, some of which are in tension with one another.

The argument of this commentary has been that politicians, policy-makers, educators, parents and the general public should be alert to the intended and unintended consequences of assessment policy decisions and initiatives, and ask whether the policies are truly fit for purpose.

Justification of the purposes that are being pursued, and of the uses to which assessment results are put, is of paramount importance. The policies must, in the end, serve to advance the education of young people.

At the heart of the matter are questions about the quality of assessment practices. We have set out some criteria for judging the quality of formative assessment, summative assessment by teachers, external tests and examinations, and accountability systems more generally.

But meeting these criteria presents challenges and we have noted what these are for: extending best practice; the professional development of teachers; enhancing dependability of assessment results; controlling costs; and creating more intelligent accountability systems.

We have not attempted to propose alternative models. Nor have we attempted to deal with all four national systems in the UK in detail. Instead, we have sought to bring the insights from recent research, including important TLRP research, to the questions that those who design systems, and the public who vote for those who make such decisions, should ask.

In the spirit of enhancing public understanding for wise decision-making we have raised some crucial questions that we believe all citizens in a democracy should ask of any assessment system.

About Warwick Mansell

Warwick Mansell is a freelance education journalist who has written about assessment policy since 2003. He was a reporter for the Times Educational Supplement from 1999 to 2008. He is the author of *Education by Numbers: the Tyranny of Testing*, published in 2007 by Politico's.

About Mary James

Mary James is part-time professor at the University of Cambridge Faculty of Education having recently retired from a Chair at the Institute of Education, University of London. She was Deputy Director of TLRP from 2002 to 2006 and directed one of its largest projects – *'Learning how to learn in classrooms, schools and networks'*. In 2007/8 she held an ESRC Programme Director's Fellowship to add value to the school based work within TLRP. She has also been a member of the Assessment Reform Group since 1992.

About the Assessment Reform Group

The aim of the Assessment Reform Group (ARG) has been to ensure that assessment policy and practice at all levels take account of relevant research evidence. In pursuit of this aim, the main targets for the Group's activities have been policy-makers in governments, and their agencies, across the UK. It has also worked closely with teachers, teacher organisations and local authorities to advance understanding of the roles, purposes and impacts of assessment.

As a voluntary group of researchers, the ARG originated in 1989 as the Policy Task Group on Assessment set up by the British Educational Research Association (BERA). In 1996, when BERA ceased to support policy task groups, the Group adopted the name ARG and its activities were funded by grants from the Nuffield Foundation. Since 2001, individual members of the Group have also been closely involved with TLRP as members of the Directors' Team and Steering Committee, and as researchers directly involved in TLRP projects and thematic work. This commentary is another contribution.

An account of how the Group has attempted to mediate between research and policy can be found in the article: Daugherty, R. (2007) Mediating academic research: the Assessment Reform Group experience, *Research Papers in Education* 22 (2): 139-153.

Members of the ARG in 2009 are: Jo-Anne Baird (University of Bristol), Paul Black (King's College London), Richard Daugherty (University of Cardiff), Kathryn Ecclestone (Birmingham University), John Gardner (Queen's University Belfast), Wynne Harlen (University of Bristol), Louise Hayward (University of Glasgow), Mary James (University of Cambridge), Paul Newton (Cambridge Assessment), Gordon Stobart (Institute of Education London).

Website for more details and publications: http://www.assessment-reform-group.org/

About this publication

This is the thirteenth in a series of TLRP Commentaries designed to make research-informed contributions to contemporary discussion of issues, initiatives or events in UK education. They are under the research programme's editorial control, but their production and distribution may be supported by sponsors. Commentaries are available from the TLRP office or at our website.

About the Teaching and Learning Research Programme

The Teaching and Learning Research Programme (TLRP) is the UK's largest investment in education research. It aims to enhance outcomes for learners in all educational sectors across the UK. Managed by the Economic and Social Research Council (ESRC) it runs, in its generic phase, to 2009, with an extension on Technology Enhanced Learning to 2012. Some 700 researchers are involved in 90 specific projects, and further work is being undertaken on the identification and analysis of common, empirically grounded themes.

About the Economic and Social Research Council

The Economic and Social Research Council is the UK's leading research and training agency addressing economic and social concerns. We aim to provide high-quality research on issues of importance to business, the public sector and government. The issues considered include economic competitiveness, the effectiveness of public services and policy, and our quality of life. The ESRC is an independent organisation, established by Royal Charter in 1965, and funded mainly by the Government.

TLRP
Institute of Education
University of London
20 Bedford Way
London
WC1H 0AL

Tel: 020 7911 5577
Email: tlrp@ioe.ac.uk
ISBN: 978-0-85473-892-2

www.tlrp.org

ARG
assessment reform group

T·L·R·P
TEACHING & LEARNING RESEARCH PROGRAMME

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL