Perspectives on Pupil Assessment

A paper presented to the GTC conference
New Relationships: Teaching, Learning and Accountability
London, 29 November 2004

GTC
General Teaching Council for England

Contents

Contributors

Carol Adams is Chief Executive of the General Teaching Council for England.

Kathy Baker is a Policy Advisor with the General Teaching Council for England.

David Hopkins is Chief Advisor to the Secretary of State for Education and Skills. Prof Hopkins is writing in a personal capacity.

Paul Black is Emeritus Professor of Science Education at King's College London.

Mary James is Reader in Education, University of Cambridge.

Caroline Gipps is Deputy Vice-Chancellor at Kingston University. Professor Gipps is also a Member of the General Teaching Council for England but is writing here in a personal capacity.

Gordon Stobart is Reader in Education at the University of London Institute of Education.

Wynne Harlen is Visiting Professor at the University of Bristol.

Editorial

Carol Adams and Kathy Baker

As the professional body for teaching, the General Teaching Council for England (GTC) is concerned with all issues related to teachers and their role in the enterprise of teaching and learning. Pupil assessment is an integral component of the teaching process and is critical to effective learning.
The Council is committed to the principle of teacher professional judgement being used to better effect in the assessment system than is the case in the current arrangements. The GTC is a public body and is therefore concerned that the assessment model for the future involves robust and transparent processes that can withstand public scrutiny.

In Autumn 2003, the GTC Council set itself the challenging task of developing recommendations on a future pupil assessment framework that would involve a greater degree of teacher professionalism as well as ensuring public accountability.
It undertook this challenge in relation to its statutory remit to provide policy advice to the Secretary of State.

In order to do this GTC staff and Members established a dialogue with the education community on pupil assessment issues during 2003-04. This took the form of a series of policy/research seminars involving key agencies particularly, the DfES and QCA, researchers, some of which are members of the Assessment Reform Group (ARG), teachers, parents, governors, LEAs and other key stakeholders. The seminar were on three themes:

- an overview of policy/research developments;
- the role of teachers in pupil assessment;
- the future balance of internal and external assessment.

The GTC produced three discussion papers that framed these seminars and their themes, the third of which formed the basis of formal advice submitted to the Secretary of State in September 2004.

The authors included in this publication made valuable contributions to the seminars and their papers provide a flavour of their original presentations. The papers are organised to reflect the order of the seminar themes and to give an overall sense of a debate in progress, in which the GTC has been privileged to participate, both as a broker and as a contributor.

The GTC's consultation was also informed by valuable discussions with teachers at teacher meetings held by the Council in Bradford and Sheffield in 2003, with discussion and e-mail consultation with teachers and stakeholders during and between seminars. Discussion and decision by the 64 teachers and other education stakeholders represented on the GTC's Council was also a key part in the process. Those processes are reflected in the content of the three GTC papers.

The GTC's recommendations on future pupil assessment arrangements frame the rationale in the third of those papers. They represent the GTC's conclusions in the light of what was heard during the consultation and what the GTC felt was achievable in the political context at the time.
Its recommendations are not intended to be set in stone but will continue to be developed and refined in response to further policy and research activity.

The Council continues to believe that teachers should have a greater role in assessment with more public credence given to their professional judgement. For that reason it recommends that the Government should invest further in Assessment for Learning to develop teachers' formative assessment skills including assessment being a greater part of the professional standards framework, advocated by Mary James in her paper. The GTC also recommended the creation of 'assessment communities' in schools and LEAs through the national strategies as described by David Hopkins, and in the development of specialist assessment roles in line with the current Chartered Examiner initiative.

However, the GTC does not believe, on the basis of the evidence that it has heard, that the majority of teachers would be able or want to have overall responsibility for summative as well as formative assessment via assessment for learning at this present time. Wynne Harlen's paper is also a reminder of some of the issues of using formative assessment for summative purposes as revealed through research. The GTC does advocate that the teacher should hold the ring between the two and have a bank of externally developed (and externally marked if required) tasks to build up summative information on individual pupils during the Key Stage. This could be fed into end of Key Stage learning related decisions, ensuring that summative assessment provides a complementary source of formative information.

A further recommendation is that summative information of that kind would form the basis of a system of pupil cohort sampling, designed to maintain the need to monitor local and national performance standards. The need to split the different uses to which assessment is put was seen as critical by the Council to prevent the current distortion of assessment for the purposes of accountability away from the focus on assessment for the purpose of furthering learning.

The Council concludes that the kind of investment that it recommends into developing teachers' formative assessment skills would result in a better basis for more future involvement in summative assessment and the moderation processes that would be involved, as well for greater longer term public trust in teachers' judgements.

The GTC welcomes the commonality in issues raised and views expressed across all the papers in this publication. In David Hopkins' paper there is a real sense of Assessment for Learning moving forward as part of Government thinking and in particular the Council supports his advocacy of the importance of building a network of assessment experts and "champions" in line with the GTC's proposals. It also supports his notion that the AfL initiative must not be too prescriptive and must, as the GTC's first paper emphasised, be built on the learning goals of individual pupils rather than driven by collective performance targets.

The impact of AfL in the classroom is as impressive described in Paul Black's article as it was in his seminar presentation at the GTC seminar last November. The change in teaching/ learning culture involved is particularly striking with pupils being "*task involved*" rather than "*ego involved*" as they tend to be in test driven competition and the relationship between teachers and pupils being gradually transformed from one of "*delivery-recipient*" to one where they are "*partners in pursuit of a shared goal*". This issue of teacher/pupil relationships is also highlighted in the second GTC paper and is at the heart of the tension between the use of assessment for learning for formative and summative purposes.

The GTC welcomes many of Wynne Harlen's conclusions to her question of whether assessment by teachers can be a dependable option for summative purposes including the use of teacher professional development to address the shortcomings of TA, the development of "assessment cultures in schools "*in which assessment is discussed constructively and positively and not seen as a necessary chore*". The need to separate TA and tests and "*ceasing to judge TA in terms of how well it agrees with test scores*" is also part of the research evidence supporting the GTC's recommendation for separating the processes of formative and summative assessment.

Caroline Gipps and Gordon Stobart's paper clearly characterise the strengths of assessment for learning as the basis of the future assessment arrangements in their framework of good assessment practice, combining a focus on learning with transparency in approach and the need to provide a sense of equity for all learners:

- *using assessment that supports learning and reflection, including formatve assessment with feedback;*
- *designing assessment that is linked to clear criteria (rather than relying upon competition with others;*
- *including a range of assessment strategies so that all learners have a chance to perform well.*

Mary James' paper confirms the GTC's conclusion to its consultation that those teachers and stakeholders involved in its seminars generally felt that with pupil assessment, there is currently an "*imbalance…with internal purposes either sacrificed or made secondary to external purposes...*where the aims for the education of students are *unlikely to be well served if*…there is only *regard to external demands*".

But it is Mary's conclusion that encompasses the importance of what is said about assessment throughout this publication and represents the GTC's own views on the future of pupil assessment so cogently:

"*…the priority must surely be the promotion of Assessment for Learning. Without learning, both as process and as outcome, educational assessment –internal or external – serves no purpose*".

Carol Adams

Chief Executive,
General Teaching Council.

Overview of current assessment policy and research developments

GTC paper 1

Introduction

This paper, the first in a series of three, represents stages in a dialogue that the Council has been conducting over the last year with key stakeholders, including those teachers actively engaged in assessment practice and research. It provides an overview of current policy developments in relation to assessment and indicates the direction of some of the current research that will be explored in greater depth in the next two papers.

Background

In general terms, assessment can be defined as being '*what we do when we take stock of how a learner is progressing [or has progressed]. How we do this, and why we do it varies tremendously…*' (Swaffield and Dudley, 2002)

The National Assessment System provides a range of information to teachers, parents and pupils indicating a child's achievements or progress. It also provides local authorities, the wider community and the Government with information on assessment outcomes on a school, local area and national basis.

As a result of the 1988 Education Reform Act, all pupils have been assessed by statutory assessment tasks and tests in core subjects at the ages of seven, 11 and 14, at the end of Key Stages 1, 2 and 3 respectively. In the 1988 Act, teacher assessment (TA) was intended to be both formative and summative. Hence there has always been a SAT level and a TA level at the end of a key stage. The change over time has been a difference in emphasis between formative and summative across subjects and at particular key stages.

GCSE provides the main framework for public examinations at the end of Key Stage 4 and the current AS/A2 provision the primary basis for assessing post-16 year olds. GCSE and post-16 public examinations and the qualifications that result have tended to drive the nature of teaching and learning post-14, so that 14-19 debates around curriculum and assessment have previously focused on issues of time and 'fit', on the tension between academic and vocational qualifications and the demands on schools in managing external examination processes. The current work developing within the Tomlinson 14-19 Reform Group and the recently published final report represents a shift from the past.

At Key Stages 1-3 debates have been around how appropriate different forms of assessment are for different aged pupils and for different subjects; the effect of a national assessment model on the breadth of the curriculum, particularly for primary pupils; and the impact of national targets and performance tables on assessment processes. At all stages, however, the debates have been underpinned by concerns for the increase of workload for teachers and pupils, the impact of assessment demands on pupil motivation and the decrease in teacher ownership and professional judgement in the process.

This change to a universal testing system throughout a learner's schooling was significant. Up to 1988, a system of sample testing existed. This was overseen by the Assessment Performance Unit (APU), which was subsequently disbanded. The change represented a critical increase in the accountability of learners, teachers and schools to government concerning their performance, particularly when linked to the other components of increased accountability in the 1988 Education Reform Act and subsequent legislation in the early 1990s.

The Government's Task Group on Assessment and Testing (TGAT), set up in 1988, differentiated between the uses served by the information resulting from assessment. It defined assessment as being:

- formative, so that positive achievements may be recognised and discussed and the appropriate next steps may be planned;
- summative, for the recording of the overall achievement of a pupil in a systematic way;
- evaluative, by means of which some aspects of the work of a school, an LEA or other part of the educational service can be assessed and/or reported on;
- diagnostic, through which learning difficulties could be identified and appropriate support provided. This could also be applied to the system as a whole. (It acknowledged that there are overlaps between diagnostic and formative assessment). (TGAT, 1988).

The original conception was that teacher assessment would be the process that would predominantly, but not exclusively, provide formative assessment and diagnostic assessment. The national standardised tasks and tests would predominantly, but not exclusively, provide summative and evaluative assessment. However, there was always an implication of a degree of crossover of roles that for a variety of factors, some technical and some political, did not happen to any substantive degree.

Key past and current debates have continued to focus on the right balance of assessment for the purposes encompassed in the TGAT model at different key stages.

While the value of formative assessment carried out by teachers in relation to their individual pupils has continued to be widely recognised, summative National Curriculum tests and tasks and public examinations have carried greater weight in public judgements concerning the performance of both pupils and teachers, schools and local authorities over the last decade. This is particularly evident in relation to the parts played by national targets, performance tables and Ofsted inspections. Teachers often question the weight given to external assessment and public examinations and the value given to what they see as narrow forms of attainment that are easier to measure.

Furthermore, the current system uses the same assessment instrument for more than one purpose. National tests are able to perform diagnostic or formative functions that are very limited compared to teacher assessment. Their main use is to aggregate data for summative and value-added purposes.

The list of stakeholders involved in assessment resulting from the 1988 Act has grown to include almost everybody: learners, parents, teachers, schools, universities, employers, community, voters and government. This has resulted in a range of sometimes competing views on what the future assessment system should look like.

This issue of accommodating and managing divergent stakeholder interests is further exacerbated by a series of policy drivers and implementation pressures. Skidmore, 2002 defines these as:

- an increasing volume of assessment with the system involving more than a million children, 54,000 examiners and moderators. It also encompassed about 25 million separate test scripts with an average pupil taking 70 exams before leaving school and a doubling in the level of a school's annual exam entry fees. Factors here have been the externally-marked National Curriculum tests at 11 and 14 in addition to teacher assessment requirements resulting from the 1988 Act, the implementation of Curriculum 2000 reforms and the additional demands this placed on the assessment system and on schools in particular;
- increasing levels of participation, including 56 per cent of 16-18 year-olds in full-time education in 2001 compared to 27 per cent in 1976, and a continued government target of 50 per cent of those under 30 entering higher education by 2010. Skidmore identifies the three principal drivers of this increase in the volume of assessment as:
  - the growing centrality of the 'standards agenda' in educational policy-making;
  - the desire to diversify and broaden the range of post-16 educational choices;
  - the increasingly direct influence for individuals of exam passes to their status, access to opportunity and earning power.
- in response to the changes outlined, the radical restructuring of the qualifications 'market' including the integration of a range of smaller examination boards, both academic and vocational, into three unitary awarding bodies, Edexcel, OCR and AQA in 1998. 1997 saw the creation of a single regulator, the Qualification and Curriculum Authority (QCA). The capacity of the three awarding bodies and QCA to manage the increase in public examinations at AS/A2 and maintain public trust in the regulatory processes was at the heart of the 2002 Tomlinson Review;
- the dependence of higher education on 'the standardised performance information that schools and public examinations provide'. Skidmore concludes that while A-level was originally designed to allow schools to filter out the small group that would go on to HE, with new patterns of schooling and participation *'it is time to re-evaluate the kinds of information [that different types of] assessment can and should be expected to provide'*. This would help prevent an unhelpful focus on schools being judged primarily on helping their students to get good grades, and on A-levels from being just a points score that represents a passport to further learning and not a qualification in itself;

- radical change in the nature of public and political expectations of the assessment and qualifications system, based on the explanation that assessment has become more 'high-stakes' - the information that assessment provides being used to make decisions that have an increasing impact on the lives of learners, teachers and on the government.

Current policy developments

There are a number of current policy developments that suggest the government is prepared to review the balance of internal and external assessment demands at particular stages of five-19 education. The Primary Strategy *Excellence and Enjoyment*, published in Summer 2003, in particular acknowledges headteachers' concerns about prioritising Key Stage 1 tests over *'a teacher's overall rounded assessment of a child's progress through the year* and about teachers and schools ending up with imposed Key Stage 2 targets based on test outcomes that they do not 'own', a process that the Strategy acknowledged '*demoralises teachers and…does nothing to raise standards*'.

Two major strands of the Primary Strategy are going ahead. They are the national implementation of the 2004 pilot with Key stage 1 testing underpinning teacher assessment rather than a parallel process. The other is a target-setting process at Key Stage 2 that will begin in schools with LEA targets being set afterwards and this will now be adopted at Key Stage 3, as indicated int eh Five Year Strategy published in July 2004. However, the issue of how teacher assessment, including moderation and exemplification processes, can underpin KS1 testing to result in a valid and reliable outcome will need further exploration.

A further indication of a change of direction in Government thinking is the emphasis found in a number of current strategies on 'assessment for learning.' *Excellence and Enjoyment defines assessment for learning as enabling 'knowledge about individual children to inform the way that they are taught and learnt.*' The DfES consultation document on core principles for the education system published in April 2003 highlighted assessment for learning as a priority for the Department and made a distinction between '*tests that promote assessment for learning as well as assessment of learning.*' There is a further governmental commitment to promote assessment for learning within what is now the secondary strategy. (DfES, 2002)

The functions of assessment for learning as promoted by the government have been further defined by David Hopkins, Director of the DfES Standards and Effectiveness Unit, as:

- using data and dialogue to diagnose the student's learning need;
- providing structured and authentic feedback for target setting;
- helping teachers adapt teaching styles to individual pupil needs;
- developing the skill of self-assessment so that pupils can take charge of their own learning. (Presentation at GTC Assessment Seminar 20 November 2003).

However, government policy on pupil assessment cannot be viewed in isolation from other government policy. A further policy driver is its commitment to what the Prime Minister referred to as 'personalised learning'. This was explained by David Miliband as '*an education system where assessment, curriculum, teaching style and out-of-hours provision are all designed to discover and nurture the unique talents of every single pupil.*' (NCSL Annual Lecture, October, 2003).

This vision underpins the Government's proposals, 14-19: opportunity and excellence published in January 2003 that suggests a more diverse and flexible curriculum post-14 with a more limited statutory core and the development of more collaborative provision between schools, colleges and workplace settings. Schools and colleges will be encouraged to enter pupils for examinations when they are ready and GCSE could be regarded as more of a 'progress check' on the route to further learning rather than the terminus that it is currently.

The issue of how the development of personalised learning as currently articulated by the Government can be fully reconciled with the some of the existing components of the accountability framework such as national tests and performance tables is unclear.

In the longer term, the Government set up the 14-19 Reform Group chaired by Mike Tomlinson with a remit to look at three areas for transformational change:

- 14-19 learning programmes;
- the development of a unified framework of qualifications;
- assessment.

The Group has been clear on its view on the current burden of assessment on learners, teachers and the system in general:

'*We believe that the balance between learning time and assessment-related demands has swung too far towards the latter…The current arrangements…stretch the practical resources of learners, schools, colleges and awarding bodies, and displace other valuable educational opportunities.*'

Its conclusion in the initial stages of the Group's work was a commitment '*to reinforce the role of assessment which is based upon the professional judgement of teachers and trainers*'. Such a change would need to be supported:

'*by measures to extend the existing capacity and expertise of schools, colleges and training providers to undertake internal assessment and re-establish the credibility of such assessment as a reliable tool for judging the achievement of young people.*' (Interim Consultation Report, 2003).

Inside the Black Box (1998); Working Inside the Black Box (2002): Assessment for Learning Research

Professor Paul Black and the Kings College London team have led a key area of research around developing formative assessment as defined by TGAT with teachers over a number of years. This needs to be viewed in context of other research work in progress on assessment, particularly in relation to the other functions of assessment in the TGAT model.

The Kings College research team defines assessment for learning in the context of its work as:

'any assessment for which the first priority in its design and practice is to serve the purpose of promoting pupils' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence.'

The research team also emphasises that:

'*An assessment activity can help learning if it provides information to be used as feedback, by teachers, and by their pupils, in helping themselves and each other to modify the teaching and learning activities in which they are engaged. Such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching work to meet learning needs.*'

Inside the Black Box set out to answer three questions. The first two of these were whether there is evidence that improving formative assessment raises standards and whether there is room for improvement. The answer to both was a clear yes. The evidence related to the second question revealed that there were three main issues for improvement with teachers' practice. These were that:

- current assessment approaches in the classroom do not promote effective teaching and learning;
- marking and grading practices tend to promote competition not personal improvement;
- assessment feedback often has a negative impact, particularly on low-attaining pupils.

The third question posed by the research, whether there is evidence of how to improve formative assessment, had a far less clear answer. Though ideas for improvement existed they were not detailed enough for teachers to implement. The Assessment for Learning Research therefore went on to plan and implement a programme in which a group of teachers have been supported in developing innovative assessment practice in their classrooms. *Working Inside the Black Box* reports the findings of its work in areas of practice such as questioning, feedback through marking, peer- and self-assessment and the formative use of summative tests.

GTC teacher testament

Meetings organised by the GTC during 2003 to consult a range of teachers on assessment issues reflected considerable consensus on the principles that should underpin the assessment model.
Overall teachers supported the view that there needed to be a better balance between assessment for learning and for accountability, that the current tests are a limited measure for assessing pupils and that there needs to be emphasis on teacher professional judgement:

'*Assessment must inform future learning rather than assessing what has gone on in the past.*'

'*Teachers are teaching to the test.*'

'*Tests are essential but imperfect – we need a holistic view of what children can do. But what about public accountability?*'

'*Give assessment back to the judgement of professionals. If you have the qualifications to be a teacher, you have the professional judgement to do the job.*'

(GTC Bradford and Sheffield Teacher Meetings, July and October 2003).

The role of teachers is obviously critical in the consideration of the principles that should underpin any assessment model. This is a key theme in the other two GTC discussion papers in the collection.

Conclusion

The developments in this paper do reflect a possible policy shift in relation to assessment in different parts of the system. In the current political climate the National Curriculum is increasingly diverse at different stages with a series of developing strategies covering primary, secondary and 14-19. While they may have an assessment for learning strand and greater emphasis on the needs of the individual pupil in common, there are still differences at specific Key Stages. Target setting processes may be changed at Key Stage 2 but there are no changes being proposed to the tests. Comments by the Secretary of State reported in the press in June 2003 suggested that a greater degree of external assessment at the end of Key Stage 3 could be a possibility in the context of age 14 becoming a more significant transition point.

Despite the Government's promotion of assessment for learning there still appears to be a tension developing. It is between the Department's view of teachers using performance data as the basis of dialogue and target-setting with pupils, and the more bottom-up approach developed in the research of responding to individual learning needs with qualitative feedback. This again seems to highlight the issue of how to reconcile the purposes of assessment for learning with assessment related to wider public accountability.

References

Assessment Reform Group, (supported by the Nuffield Foundation), 2002, *Testing, Motivation and Learning*, University of Cambridge.

Black, P, Harrison, C, Lee, C, Marshall, B, William, D, 2002, *Working Inside the Black Box*, Kings College, London.

DfES, 2003, *Excellence and Enjoyment*, DfES Publications

DfES, (2003) *New Specialist System: Transforming Secondary Education* DfES Publications

Skidmore, P, (2003) *Beyond Measure*, DEMOS

Swaffield, S and Dudley, P, (2002), *Assessment Literacy for Wise Decisions*, Association of Teachers and Lecturers, London.

Working Group on 14-19 Reform, 2003, Reforming the 14-19 *Curriculum and Qualifications: Summary of progress*, DfES Publications.

Assessment for personalised learning: The quiet revolution

David Hopkins

Introduction

This is a crucial time for our education system. We need to build on the progress which has made it among the best in the world by remaining at the leading edge of change and by securing world-class standards for all. The key to meeting this challenge in the next phase of educational reform is to personalise learning so that every individual can reach his or her full potential.

Definition

Personalisation is a major theme of public service reform and is one of the five principles informing the Government's Five Year Strategy for Children and Learners [1].

Personalised learning means:

- tailoring educational provision to meet the needs and aspirations of individual learners within a social context to maximise their achievement as independent, lifelong learners;
- high expectations for all sustained through high quality teaching based on a sound knowledge and understanding of each child's needs;
- designing teaching, curriculum and school strategies to create a coherent learning system tailored to the individual pupil;
- personalising the school experience to remove barriers to achievement and bring about the best conditions for learning [2].

The rationale

In recent years there have been major achievements in the education system across the age range. In primary schools higher standards in literacy and numeracy have been sustained: we are ranked third in the world for reading, and first in English-speaking countries [3].

At Key Stage 3 there have been improvements in every subject at every level. GCSE and A Level performance continues to rise steadily. These results are achieved, as Ofsted evidence shows, in the context of teaching which is better than ever, ICT which is beginning to transform teaching and learning in the primary and secondary phases and schools which are continuing to improve in management and leadership.

Despite the undoubted successes however, significant challenges remain, particularly in narrowing achievement gaps and improving long-term participation in education. International studies (PISA 2001 [4]) show a wider gap between higher- and lower-attaining children here than elsewhere, and a stronger link between social class and achievement. This link continues and widens through the years of schooling so that by the age of 16, 75 per cent of middle class youngsters get five good GCSEs, but only 25 per cent from working class backgrounds do so. Crucially, then, we have low participation rates in educational pathways post-16.

The variation in student performance also arises from variation in quality within and between schools, which restricts opportunity and choice. Some studies show that there is four times as much variation within schools as between schools [5]. All this argues for an approach to school improvement that at the same time focuses on the organisational conditions of the school as well as the organisation of teaching and learning. Personalised learning provides the means of meeting this challenge and, by ensuring high standards for all, builds a system which produces high excellence and high equity.

Until recently, universal services have been seen as the means by which equity can be realised. But the historical association of a universal coverage with standardised provision is now under extreme pressure. In part this is because the emphasis on universalism has not meant improvements for all and for some groups of pupils has not delivered greater equity or choice.

The next phase of educational reform needs to build on the platform for progress which has been established, and to address the major challenges which remain, so that excellence and equity are combined. In order to do this, the system will need to move from standardised provision with uncontrolled variation in quality to personalised provision based on consistently high quality, where variation is controlled and actively tailored to individual pupils' needs and aspirations, to ensure that it is the achievement of full potential that becomes universal.

Assessment for learning

The most powerful lever we can pull at the moment to achieve personalised learning is assessment for learning. Assessment for learning has been defined 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.*' [6]

It is at the forefront of developing personalised learning because it is a powerful means of helping teachers to tailor their teaching to pupils to get best improvement, and to involve, motivate and help them to take the next steps in learning.

David Hopkins' views as expressed here are personal and not given on behalf of the Department for Education and Skills or the Secretary of State for school standards.

Assessment for learning:

- is embedded in a view of teaching and learning of which it is an essential part;
- involves sharing learning goals with pupils;
- aims to help pupils to know and to recognise the standards they are aiming for;
- involves pupils in [peer and] self-assessment;
- provides feedback that leads to pupils recognising their next steps and how to make them;
- is underpinned by confidence that every pupil can improve;
- involves both teacher and pupils in reviewing and reflecting on assessment data [information]. [7]

Central to assessment for learning is the focus on helping pupils become increasingly effective independent learners. Teachers need to develop a good understanding of subject progression so that they can help pupils:

- understand precisely what they are trying to learn and why, and what their next steps are;
- assess their own progress (and similarly help their peers); and,
- recognise the standards they are aiming for and strive for personal excellence.

Teachers also need to continue to develop their understanding of how pupils learn so that they can help them to:

- reflect on how they learn;
- develop learning strategies and apply them in different circumstances;
- engage in high-quality classroom dialogue with the teacher, other adults and their peers in order to develop as effective independent learners.

Assessment for learning:

- provides a framework to help structure and focus the whole-school development of teaching and learning;
- gives teachers a shared language and context within which they can develop their teaching skills, such as questioning, modelling, explaining and providing informative oral and written feedback;
- helps establish a learning environment in which the respective roles and responsibilities of pupils and teachers are better understood, pupils increasingly take responsibility for their progress and become more actively engaged.

Since assessment for learning is a key component of personalised learning, a large part of the success of personalised learning, and the fulfilment of the radical agenda for change it presents, will depend on whether high-quality assessment for learning can be developed powerfully and consistently through the education system. Past evidence, however, underlines the extent of the challenge to be faced if assessment for learning is to be implemented and sustained in practice.

Assessment for learning: a recurrent weakness

It is important to stress, of course, that the concept of assessment for learning is not new. There have been exciting developments in recent years which involve a wide range of key partners, including King's College and the Assessment Reform Group with the seminal Black Box series of research findings [8], LEAs and schools and the National Strategies.

Nevertheless, although significant gains have been made and there are examples of outstanding practice, Ofsted identifies assessment and its application to teaching and learning as comparatively weak areas [9]. Too many schools lack adequate systems for tracking the progress of individual pupils.
The challenge is especially marked for those pupils with special educational needs to ensure that their individual needs are met consistently across all subjects.

Assessment for learning: going wider and deeper

There are few schools where we could say that assessment for learning is presently well-established across all classes and teachers to reach all pupils. More and more pupils need to benefit from assessment for learning. So we need to go wider. The phrase is often used, but it is not always clear that we all mean the same thing. We need to develop the strategies and techniques, but more than this we need to construct a shared understanding nationally of what assessment for learning entails and of how it sits within teaching and learning, so that we are secure in the rationale of how and why it works. So we need to go deeper.

Assessment for learning developments

We are building on the good practice available in both phases to develop that common understanding to ensure that all pupils benefit from assessment for learning. The model of development is school-based, collaborative, whole staff enquiry.
The involvement of senior management teams is critical. How the work is co-ordinated and supported and how it links with other work to improve teaching and learning across the school, will determine the extent of its long-term impact on attainment.
The work in assessment for learning has two mutually dependent strands:

- the use of data to diagnose and target pupils' individual learning needs and challenges;
- using teaching and learning strategies to create more powerful learning experiences for pupils.

The data strand

The data strand is about using data intelligently to inform teaching and to move individuals forward. Data is not an end in itself. It helps teachers, subject leaders and the senior leadership team identify underperformance, and do something about it. In this sense it is the most valuable currency in school improvement. More than this, the data is a moral challenge to raise quality and equality in our education system. For example, the Pupil Achievement Tracker (PAT) software helps schools review their performance and analyse pupils' past and current attainment so that they can tailor lessons and progression to pupils' needs.

Data covers a range of sources of information, including formative and summative data, work samples, lesson observation and pupil opinions. It means quantitative, numerical outcomes and qualitative, curricular outcomes. Numeric targets need to be translated into meaningful curricular targets which are negotiated with and understood by pupils. The use of data becomes assessment for learning at the point where this negotiation takes place and then helps to focus learning and improve pupil outcomes.

Analysed in terms of individuals in this way, the data can be used to inform assessment for learning, but it will be equally important to discuss it in terms of broad patterns and trends for school improvement. All our partners rightly regard high-quality information as critical to the new relationship with schools. Sophisticated, user-friendly, multi-use analyses are needed by schools for self-evaluation, by school improvement partners for their "single conversation" with a school, and by inspectors. Simple, user-friendly, informative data is needed by parents in the Profile.

We need to ensure that data at this level is transparent, clearly understood by practitioners and the public and of high quality. We also need to ensure headteachers and teachers have the willingness to discuss it openly and can develop teaching and school improvement plans based on it. Again, the PAT will be a valuable tool in improving such 'assessment literacy' and having collective ownership of such information is a key indicator of a professional learning community.

The teaching and learning strand: The contribution of the National Strategies

The Primary and Secondary National Strategies are the key delivery platforms for the teaching and learning strand. They are undertaking the largest-ever initiative (both nationally and internationally) to support the development of assessment for learning in schools.

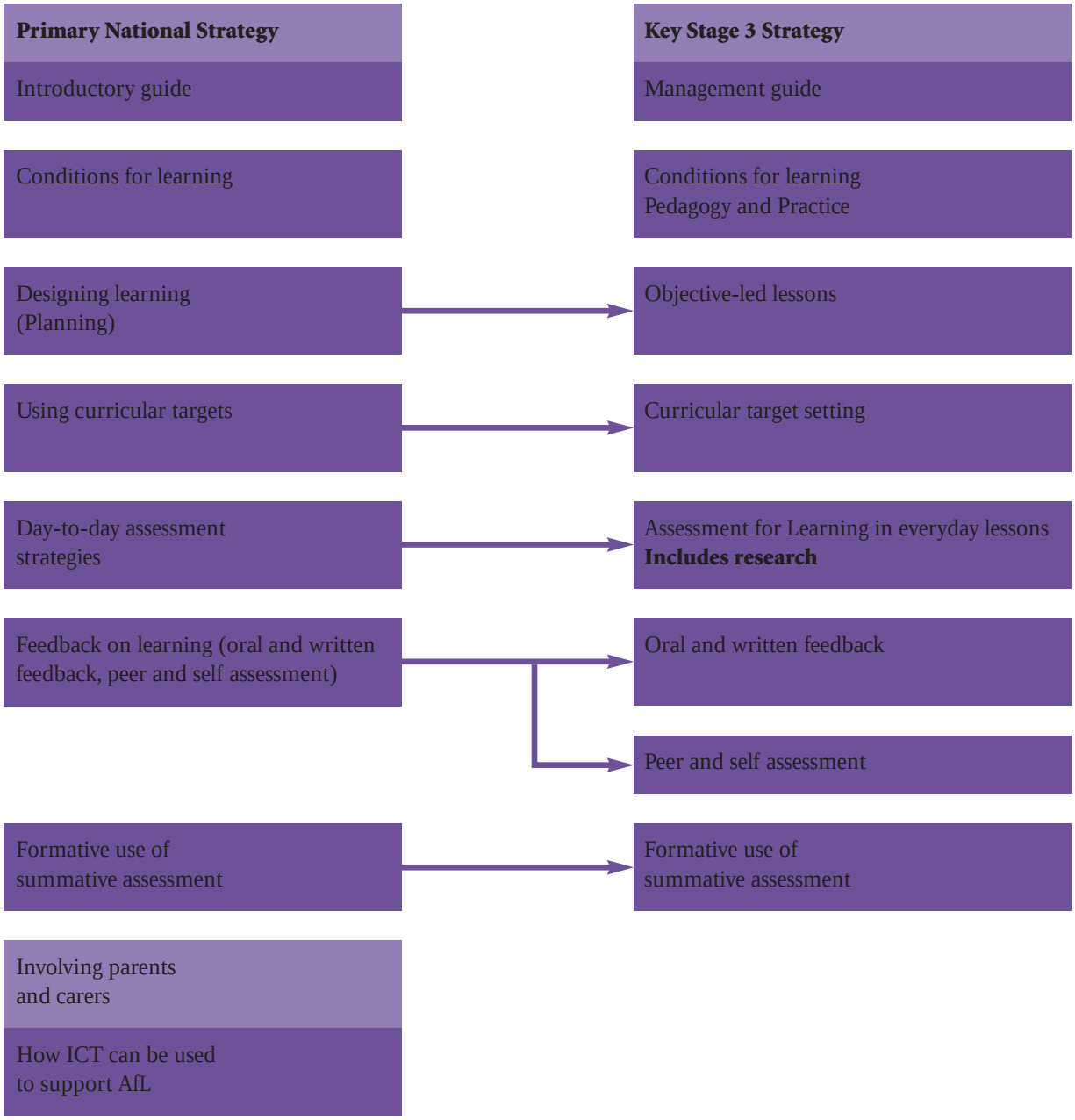
Assessment for learning is integral to the new primary resource *Excellence and Enjoyment, Learning and Teaching in the primary years*[10]. A management guide with a self-evaluation grid and a series of training units have been written based around three themes: creating a learning culture; understanding how learning develops; and planning and assessing learning. It is a key element of the Primary Leadership Programme. Primary consultants are trained to support schools' implementation of the Learning and Teaching resource CPD materials and the whole school implementation of assessment for learning.

In the Secondary Strategy (formerly the Key Stage 3 Strategy), a management guide and generic training materials in Assessment for Learning have been produced[11] accompanied by subject development materials for 12 subjects (English, mathematics, science, ICT, geography, history, design and technology, modern foreign languages, art and design, music, physical education and religious education). All secondary schools receive a core training day. Those who opt to make it a whole-school development priority receive a further £1,000 and four days' consultancy support. A self-study unit for *Pedagogy and Practice: teaching and learning in secondary schools*[12] is currently available.

Additional generic training materials on key aspects of assessment for learning, dialogue and questioning and securing progression are being piloted as part of a research project; guidance on coaching is also being produced, plus support materials for teaching assistants, all of which will be simultaneously available in 2005.

The AfL work undertaken by the Primary and Key Stage 3 Strategies is closely co-ordinated. The following chart shows how the training materials link together across the phases.

Assessment for learning progression: Primary strategy to key stage three strategy



That both Strategies have been working closely together on this development is vital since it has critical relevance to transfer and transition. If pupils are used to the kind of working involved in assessment for learning, then this will open up the way to enhance continuity and to ensure that the progress pupils make is maintained.

Assessment for learning will be included in materials for training new headteachers and in Initial Teacher Training to build capacity. In addition to this we will publish guidance on how assessment for learning can be particularly beneficial for pupils for whom English is an additional language. All LEAs have appointed lead consultants for assessment for learning and many have included a commitment to it in their education plans.

The importance of dialogue

Much work is taking place, but the development of assessment for learning across all settings and classrooms is a process which is going to take time. The support for it will be sustained. It involves training and consultancy, but moving beyond the training to make it work in all classrooms consistently and powerfully will be the main challenge and secure the greatest impact. This requires sustained dialogue. It is not about imposing on the system. Having the thinking and tools to do the job frees teachers to act as informed, creative professionals.

Consolidating and extending work in assessment for learning will mean:

- strengthening the links between Strategy teams and other LEA teams and services;
- running networks of assessment co-ordinators and consultants so that they are supported to be assessment for learning champions, building on the two conferences for assessment advisers held in spring 2004 and the ongoing networking of strategy consultants;
- assisting schools in their development of policy, thinking and approaches to assessment for learning;
- the sharing of best practice;
- sustaining improvement and ensuring there is a clear impact on the classroom, by ensuring that assessment for learning informs schools' use of coaching and networking.

In this way we will enable skilled practitioners to make pupils partners in learning, to help them judge their own work, to reflect on how they and others learn and to set and achieve future learning goals.

Conclusion

Although assessment for learning is about raising standards of learning and achievement, it is also about more than this. Through building ownership of the teaching and learning process among learners and teachers, it offers the opportunity for a radical redefinition in the culture of classroom practice. Such a fundamental shift will be brought about by giving teachers the opportunities and tools to create ever more powerful learning experiences and by helping pupils to reflect on and take control of their own learning. This is the quiet classroom revolution.

David Miliband in his North of England speech, said: '*Giving every single child the chance to be the best they can be, whatever their talent or background, is not the betrayal of excellence; it is the fulfilment of it*'. (Personalised learning: building a new relationship with schools, David Miliband, Minister of State for School Standards, North of England Education Conference, Belfast, 8th January 2004.)

Every child is special and creating an education system which treats them so is what personalised learning is about. That means overcoming the false dichotomies and the either/or which have bedevilled the education system, so that for all pupils learning means not either/or, but both/and, both excellence and enjoyment, skills and enrichment, support and challenge, high standards and high equity, breaking the link between socio-economic disadvantage and attainment, present success and long-term participation, deep engagement and broad horizons. That is our goal for personalised learning and going wider and deeper with assessment for learning will be a principal means of achieving it.

References

1. Five Year Strategy for Children and Learners (Para. 2, 'Definition', p.1) Department for Education and Skills (2004), Department for Education and Skills Five Year Strategy for Children and Learners, London, TSO. Cm 6272. ISBN 010162722X

2. For more information about personalised learning, see DfES pamphlet 'A National Conversation About Personalised Learning' DfES/0919/2004 which is downloadable and available for online ordering from **www.teachernet.gov.uk** and the website **www.standards.dfes.gov.uk/personalised learning**.

3. Reading All Over the World2001: Progress in International Reading Literacy Study (PIRLS). England came third in the study of reading achievements of 140,000 10-year-olds in 35 countries, with only Sweden and the Netherlands getting better results. Two test papers were taken by each of 140,000 10-year-olds. The papers were divided into short stories and non-fiction pieces, each 400-700 words long. Children were either given one of each type, or two of the same type. The full report can be downloaded from **http://www.dfes.gov.uk/research**

4. OECD (2001) Knowledge and Skills for Life: First results from PISA 2000 Appendix B1, Table 8.1, p.308 PISA 2000. The OECD programme for international student assessment (PISA) is a three-yearly survey (PISA 2000, PISA 2003, PISA 2006...) of the knowledge and skills of 15-year-olds in the principal industrialised countries. It assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society. **http://nces.ed.gov/surveys/pisa/**

5. OECD (2001) Knowledge and skills for life: First results from PISA 2000, Fig. 2.6, p.61

6. Assessment for Learning: 10 principles, Assessment Reform Group, 2002.

7. Assessment for Learning – Beyond the Black Box, Assessment Reform Group, 1999 University of Cambridge, School of Education.

8. Black, P and William, D (1998), Inside the Black Box: Raising Standards Through Classroom Assessment, London, King's College London. ISBN: 1871984688. Assessment Reform Group (1999), Assessment for Learning: Beyond the Black Box, University of Cambridge, School of Education. Working inside the Black Box: assessment for learning in the classroom. Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2002). Working inside the Black Box: assessment for learning in the classroom. London, UK: nferNelson. ISBN 1 871984 39 4. **http://www.assessment-reform group.org.uk/publications.html**

9. Standards and Quality 2002/2003 - The Annual Report of Her Majesty's Chief Inspector of Schools 4 February 2004 0-10-292677-8 and Summary of the Annual Report of Her Majesty's Chief Inspector of Schools p.5, p. 10 and p. 12. See also Good Assessment in Secondary Schools 2003 HMI 462, March 2003 and Good Assessment Practice in... series of 13 subject documents, March 2003

10. Excellence and Enjoyment: Learning and Teaching in the primary years DfES 0518 2004 G.

The Management guide for the Excellence and Enjoyment: Learning and Teaching in the Primary Years: Introductory guide: supporting school improvement was sent to all schools in May 04 DfES 0344-2004 G.

11. A management guide and generic training materials on assessment for learning have been produced for Key Stage 3 and sent to all schools: Assessment for learning: whole school training materials DfES 0043 2004 G.

12. Pedagogy and Practice: Teaching and learning in secondary schools, DfES 0423 2004 G.

David Hopkins is writing in a personal capacity.

Raising standards through formative assessment

Paul Black

The background

The ideas I will discuss have their origin in a review of research published in 1998 (see Black & Wiliam 1998). This work established that there was strong evidence that formative assessment can raise standards of pupil achievement, but that the assessment practices entailed were not implemented in most classrooms. The group at King's College went on to explore the potential for practical improvement by collaborating with a group of teachers willing to take on the risks and extra work involved, with support from their schools and LEAs. Through collaboration with Medway and Oxfordshire LEAs we were able to recruit six secondary schools spanning a range of catchment backgrounds. Initially, 12 science and 12 mathematics teachers were involved; later on, 12 teachers of English joined in the work.

Almost all the teachers were positive about the value of the project and there were significant gains in test performance for the classes involved. We summarised our findings in a second short booklet for teachers (Black et al. 2002), and reported them at length in our book on Assessment for Learning (Black al. 2003). The development of formative assessment has since been made a significant component of the DfES initiative for Key Stage 3.

Four main activities

As we have tried to summarise the results of research, and of the experiences of teachers, four ways to implement formative assessment have emerged. What they all have in common is the focus on feedback, from student to teacher so that the teacher can understand the learning needs of the students, and from teacher to student whereby the teacher adjusts her contribution to meet these learning needs. This continuous adaptation of teaching in the light of communication of students' thinking is the key to formative assessment, ie assessment that promotes learning.

Questioning and classroom dialogue

The first of the four ways is concerned with the to-and-fro of discussion in the classroom. Here is how one teacher summed up her experience:

- "*My whole teaching style has become more interactive. Instead of showing how to find solutions, a question is asked and pupils given time to explore answers together. My Year 8 target class is now well-used to this way of working. I find myself using this method more and more with other groups.*
- "*Unless specifically asked pupils know not to put their hands up if they know the answer to a question. All pupils are expected to be able to answer at any time even if it is 'I don't know'.*
- "*Pupils are comfortable with giving a wrong answer. They know that these can be as useful as correct ones. They are happy for other pupils to help explore their wrong answers further.*" - Nancy, Riverside School.

This teacher had realised that her questions had to help students to express their ideas about the important concepts to be learned and, that if students are to give an honest account of their ideas, the habit of competing to give the right answer has to be changed. She also came to allow more time between asking a question and expecting an answer ('wait time') in order to encourage pupils to think, and to share ideas with one another. What she does not mention is the difficult skill that she and others had to develop, of responding on one's feet to whatever comes up so as to explore or challenge in order to develop students' thinking.

The only point of asking questions is to raise issues about which the teacher needs information or about which the pupils need to think. Where such changes have been made, experience has shown that pupils come to realise that learning may depend less on their capacity to spot the right answer and more on their readiness to express and discuss their own understanding. The teachers also shift in their role, from presenters of content to leaders of an exploration and development of ideas in which all pupils are involved.

Comment-only marking

Research evidence has shown that, whilst giving comments on pupils' written work can improve learning, giving then marks or marks with comments produces hardly any improvement. Teachers can understand this, because marks alone do not give any guidance about how to improve. The key principle is that, as with oral questioning, written tasks should encourage pupils to develop and show understanding of the key features of what they have learnt. When pupils focus on marks this encouragement is undermined. So there should be no marks, and, moreover, comments have to be carefully formulated to identify both what has been done well and what still needs improvement. They should give guidance on how to make that improvement. Opportunities for pupils to follow up comments should be planned as part of the overall learning process.

The central point is that, to be effective, feedback should cause thinking to take place. Implementation of such practice can change the attitudes of both teachers and pupils to written work: the assessment of pupils' work will be seen less as a competitive and summative judgement and more as a distinctive step in the process of learning.

Peer- and self-assessment

Engaging in peer- and self-assessment is much more than just checking for errors or weaknesses. It involves pupils in explaining their work to their peers and in listening to their comments and reactions. It thus involves pupils in being clearer about what they are trying to achieve, as well as requiring them to be active in their learning. As one pupil wrote: '*After a pupil marking my investigation, I can now acknowledge my mistakes easier. I hope that it is not just me who learnt from the investigation but the pupil who marked it also*'.

Our experience of work on this theme shows the first important step is that the criteria for evaluating any learning achievements must be made transparent to pupils. This will enable them to have a clear overview both of the aims of their work and of what it means to complete it successfully. Such criteria may well be abstract - concrete examples should be used in modelling exercises to develop understanding.

For self-assessment a useful guide is to ask pupils to 'traffic-light' an end-of-topic test in the first lesson on the topic: the amber and red items can be used to re-adjust priorities within the teaching plan. In order for peer-assessment to be effective, pupils must be taught the habits and skills of collaboration in peer-group working, both because these are of intrinsic value and because peer-assessment can help develop the objectivity required for effective self-assessment. The main point here is that peer and self-assessment make unique contributions to the development of pupils' learning - they secure aims that cannot be achieved in any other way.

Formative use of tests

A first step in any attempt to make summative tests helpful for learning is that pupils should be engaged in a reflective review of the work they have done to enable them to plan their revision effectively. Peer- and self-assessment practices are very helpful here. A second step is to ask pupils to write some questions as this activity calls for, and so develops, an overview of the topic: '*Pupils have had to think about what makes a good question for a test and in doing so need to have a clear understanding of the subject material. As a development of this, the best questions have been used for class tests. In this way the pupils can see that their work is valued and I can make an assessment of the progress made in these areas.*' Angela, Cornbury Estate School.

A final step is to involve pupils in marking their own and one another's work. To do this they have to develop, and/or apply, criteria of quality to the test responses, and as they understand these criteria better they can begin to understand how their work might be improved. Then it may be worthwhile to provide opportunities for pupils to rework examination answers in class.

The main overall message is that summative tests should be, and should be seen to be, a positive part of the learning process. By active involvement in the test process, pupils can see that they can be beneficiaries rather than victims of testing, because tests can help them improve their learning: '*They feel that the pressure to succeed in tests is being replaced by the need to understand the work that has been covered and the test is just an assessment along the way of what needs more work and what seems to be fine.*' Belinda, Cornbury Estate School.

Summary

Learning principles

Two of the main ideas developed by research are that learning work should start from a learner's existing understanding and that the learner should be actively involved in the learning process. The practices described above are productive because they exemplify these principles, and also because they put into effect two other research findings. They are that teachers should develop the learner's overview, ie meta-cognition, which requires a view of purpose, understanding of criteria of quality of achievement, and self-assessment; and that they should develop social learning, ie learning through discussion with one's peers (Wood, 1998).

Self-esteem

Learning is not just a cognitive exercise: it involves the whole person. The need to motivate pupils is evident, but it is often assumed that this is best done by offering such extrinsic rewards as merits, grades, gold stars and prizes. There is ample evidence that challenges this assumption. Pupils will only invest effort in a task if they believe that they can achieve something. If a learning exercise is seen as a competition, then everyone is aware that there will be losers as well as winners: those who have a track record as losers will see little point in trying.

Thus, the problem is to motivate everyone even though some are bound to achieve less than others. In tackling this problem, the type of feedback given is very important. Many research studies support this assertion (Dweck 2000).

In general, feedback given as rewards or grades enhances 'ego-involvement' rather than 'task-involvement'. It can focus pupils' attention on their 'ability' rather than on the importance of effort, damaging the self-esteem of low attainers, and leading some high-attainers to avoid any task that challenges because of fear that they will turn out to be not so bright after all. Feedback that focuses on what needs to be done can encourage all to believe that they can improve. Such feedback can enhance learning, both directly through the effort that can ensue, and indirectly by supporting the motivation to invest such effort: *Here the classroom ceased to be a habitat where only the brightest survived and flourished, but one where, with careful grouping and good questioning, every student could feel themselves making progress through the lesson.* (Higton p.90 in Black et al. 2003)

Teacher change and pupil change

One of the striking features of the project was the way in which, in the early stages, many spoke about the new approach as 'scary', because they felt that they were going to lose control of their classes. Towards the end of the project, they described this same process not as a loss of control, but as one of sharing responsibility for the class's learning with the class - exactly the same process, but viewed from two very different perspectives. In one perspective, the teachers and pupils are in a delivery-recipient relationship, in the other they are partners in pursuit of a shared goal: '*What formative assessment has done for me is made me focus less on myself but more on the children. I have had the confidence to empower the students to take it forward.*' Robert, Two Bishops' School

References.

Black, P. & Wiliam, D. (1998) Inside the Black Box: raising standards through classroom assessment London: nferNelson. See also Phi Delta Kappan, 80(2), 139-148.

Black, P. J.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D (2002) Working inside the Black Box: Assessment for learning in the classroom. London: nferNelson.

Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2003). Assessment for Learning: putting it into practice. Buckingham, UK: Open University Press.

Dweck, C. S. (2000) Self Theories: their role in motivation, personality and development. London: Taylor and Francis.

Wood, D. (1988) How children think and learn. 2nd.edn. Oxford: Blackwell.

The role of the teacher in pupil assessment

GTC paper 2

Introduction

Since 1988, teachers have played a key role in formal pupil assessment as part of the National Curriculum Framework. In the curriculum core subjects, teachers have been involved in conducting external tests at the end of Key Stages 1, 2 and 3. They have also been responsible for teacher assessment summative judgements at the end of those three Key Stages, to sit alongside the test results in the core subjects and to stand alone in subjects outside the core. At Key Stage 4 and post-16, teachers have been responsible for assessing varying percentages of GCSE, A-level and GNVQ coursework/assignments, as well as being involved in the preparation and invigilation related to pupils sitting public examinations. Over time, a decreasing number of teachers have been employed externally by awarding bodies to be involved in the marking and moderation processes related to public examinations. Right through the system, teachers have additionally continued to assess pupils formatively on an everyday basis in the classroom. The distinction between summative and formative assessment follows the Task Group on Assessment and Testing (TGAT) definitions as included in the first GTC paper in this series.

This second GTC discussion paper focuses on the role of the teacher in the current pupil assessment arrangements. It also reviews the messages that emerge from research about the kind of support that teachers will need for a broader and more effective role in the future assessment of pupils.

The role of teachers since 1988

As the first GTC paper indicated, summative assessment outcomes reached by the teacher, though continuing to be professionally respected, have carried less public weight than the outcomes of the end-of Key Stage Tests. It is test and public examination outcomes, with related value-added data for example, which dominate published performance tables. As Singh emphasises, the TGAT report originally tried to marry the needs of the Government's agenda for a national framework to compare school and LEA performance with those of teachers in supporting teaching and learning. However, as the assessment arrangements were made more manageable, its '*instruction supporting aspects gave way to accountability concerns*' (Singh, 1999) One significant outcome for teachers of the Dearing Review of the National Curriculum and Assessment in 1993-5 was that LEAs were no longer required to provide moderation for teacher assessment. This was a response to teacher workload concerns but also represented a reduction in the importance attached to teachers arriving at summative assessment judgements via collective professional learning processes.

Broadfoot, 1999, characterises three broad themes as '*connecting narratives in the history of assessment policy in England during the last 25 years*'. These are 'performativity', with assessment used for measurement and accountability purposes; 'empowerment', characterising various initiatives which cast assessment in the very different role of supporting learning, including the Black/Wiliam led *Assessment for Learning* action research; and 'certification', '*including the attestation of competency and selection – the more familiar territory of assessment purposes*'. The tensions between these three assessment purposes and interlinking strands of current policy debate encompass the challenges faced by teachers in their core role of assessing the pupils they teach.

Teachers and tests

The impact of the changes to the teacher role brought about by the requirement for external testing of pupils was perhaps greater in primary schools than in secondary schools where teachers had been used to teaching towards public examinations and using department/school testing as summative instruments. A survey and telephone interviews carried out by the Institute of Education in 1995 with Year 2 and Year 6 teachers and headteachers reflected professional concerns about the effects of the tests that teachers were conducting on teaching and learning and teachers' own assessments. These included that:

- some Key Stage 1 teachers identified that they lost up to a half term in teaching and learning; 75 per cent of Key Stage 1 and two thirds of the Key Stage 2 teachers in the study were concerned about the levels of distress of some particular children towards the tests;
- some Key Stage 1 teachers at the time of the survey claimed that they had not prepared children for the tests, but two thirds said that they would do more practice tests in the future. At Key Stage 2 the change was more marked with a period of revision in the summer term and an emphasis on the core subjects. The research concludes that '*teachers will teach to the tests if the stakes are high*' (Clarke, 1995);

- the majority of Key Stage 1 teachers in the study felt that the quality of their teacher assessment was enhanced by the tests/tasks whereas 50 per cent of Key Stage 2 teachers claimed to be unaffected. (This may have been due to the different timescales that the tests had been running at the time);
- overall, 85 per cent of Key Stage 1 teachers and 83 per cent of Key Stage 2 teachers felt that continuous teacher assessment is more useful to teachers, pupils and parents than the results of external tests and tasks.

The response of the Government to the concerns about teacher workload related to the Key Stage 1 tests and tasks in particular was to move towards more 'paper and pencil tests' and less integration of the tests into the normal classroom.

Statutory teacher assessment

Gipps/Clarke, 2000, define statutory Teacher Assessment (TA) as being where:

'*teachers make an assessment of each pupil's level of attainment on the scale of levels in relation to the attainment targets…Teachers make these assessment in any way they wish, but observation, regular informal assessment and keeping examples of work is encouraged*'.

The researchers also highlight the significance to TA of the Dearing Review of the National Curriculum and assessment arrangements set up in 1993, which shifted away from the almost 1,000 original multiple statements of attainment to around 200 broad level descriptions. End-of-Key Stage TA level judgements from then on were to be based on the level descriptions from the eight-level scale of the Attainment targets for the various subjects and were to be arrived at through a 'best fit' approach.

Gipps/Clarke carried out three research projects in 1996-8 for the then School Curriculum and Assessment Authority (SCAA), the first of which in 1996 was to monitor the consistency of TA in England across Key Stages 1, 2 and 3 (Gipps and Clarke, 1996) The project involved questionnaires and visits and included Year 2, Year 6 teachers, assessment co-ordinators and heads of core subject departments in secondary schools. The project looked at three dimensions of TA judgements:

- ongoing day-to-day assessment judgements;
- end-of-Key Stage summative level judgements;
- whole-school or department standardisation meetings.

The findings reflect different teacher roles in assessment across subjects and across phase. In terms of ongoing assessment judgements, Key Stage 3 maths and science teachers used formal approaches to assessment such as end of module/classroom tests whereas teachers of English at Key Stage 3 and primary teachers used pupil self-assessment and pupil portfolio evidence.

In relation to evidence used for end-of-Key Stage summative-level judgements, primary teachers and heads of English departments were more likely to consider dialogue with the pupil as a source of information, and primary teachers were more likely to rely on memory, whereas secondary teachers more often used homework as a source of evidence. In terms of the use 'best fit', findings showed that teachers mostly found it too vague but welcomed its increased manageability. In using level descriptions to arrive at a summative level, secondary teachers were more likely to average the set of levels which pupils had at the end of the year while primary teachers used a variety of 'best fit' approaches.

On moderation all schools involved found meetings useful. Secondary schools appeared to use meetings to check on marks awarded for school-based tasks or tests with maths and science using them to devise and moderate the assessment of Attainment Target 1 tasks and English departments moderating samples of pupil's writing. By contrast, primary schools used a range of pupils' ongoing work as the focus for meetings and for arriving at a school definition of a level:

'*Arriving at a standardised interpretation of levels was used by primary teachers for more than deciding final TA levels; having a clear view of progression from level to level for each attainment target helped teachers plan work which was appropriate for their age group*'.

It is difficult to generalise from such a complex picture. But secondary maths and science departments in particular seemed to be adapting the TA requirements into their existing formal approaches to assessment. Primary teachers seemed to be using TA in a variety of ways but, in some cases, as a means of re-defining their teaching and learning and pupil progress at key points.

Teachers' perceptions of assessment purposes

Singh's research, (1999) involved interviews with Key Stage 2 and 3 teachers teaching science. Teachers were asked about their definitions and use of formative assessment and about the distinction they saw between formative and summative assessment. The findings concerned a confusion in teacher's views of the distinction between formative and summative assessment; teachers in general were sympathetic to formative assessment but in some cases saw it as involving unrecorded informal processes. Some thought they were using formative assessment most of the time, and some teachers did not see much difference between formative and summative. In fact, much of what the teachers described to Singh were summative processes around grouping and setting pupils and she concluded that many teachers had gained expertise in carrying out summative assessment and had adapted processes imposed on them into their own teaching style. However the most revealing finding overall was teachers' lack of clarity about the purposes of assessment.

'Confident judges' versus 'Measurers'

Research by Hall and Harding in 1999 supported the Gipps/Clarke study in findings about teachers employing a variety of approaches to the 'best fit' requirement for TA with some teachers looking for an exact as opposed to a best-fit approach. The research involved six primary schools in six different LEAs in the north-east of England and sought views from teachers, assessment co-ordinators and LEA officers. The study found that most teachers were confident about the reliability and validity of their assessment assessors on an everyday basis, but some became less confident when faced with the task of formally allocating levels at the end of Key Stage 1. It identified two main categories of teacher role in the process defined partly by the nature of school in which they taught. 'Confident judges' tended to work in schools where collegiality, collaboration and moderation procedures were fully embedded in the culture of the school. 'Measurers' were less secure, more isolated and more reliant on procedures. They were more likely to break the level descriptions back down into quasi-statements of attainment in order to arrive at what they saw as a more secure judgement.

Further research carried out by Hall and Harding, 2002 on level descriptions and teacher assessment, again involving six schools in six different LEAs supports the earlier findings about individual teachers' TA confidence corresponding to the extent to which schools are 'assessment communities'. Four out of the six school made TA the business of the whole staff where processes like moderation and interpretation of level descriptions were shared, while in the other two schools, TA had been placed 'on the back burner'.

Gipp's conclusion on the lack of consistency in teachers in interpreting TA 'best-fit' in particular is that it is 'not acceptable in a high-stakes programme'. The issue of enhancing teacher assessment is re-visited later in this discussion.

A further issue from the Harding/Hall research is their finding that teachers avoid incorporating non-curriculum factors into their level decision-making as demanded by the current assessment model. The researchers conclude with the question of how, if at all, motivational and disposition to learn should be '*recorded, reported and used*'. These are issues that are at the heart of Assessment for Learning thinking.

Assessment for Learning (AfL)

As the first paper in this series made clear, the first priority of the assessment for learning research is to promote student learning. Those researchers involved in the work see formal tests as being '*isolated from normal teaching and learning, carried out on special occasions with formal rituals and often conducted by methods over which teachers have little or no control*'. (Black et al, 2003) AfL promotes formative assessment by developing the use of evidence methods to adapt teaching work to meet learning needs.

AfL research principles include that it:

- is part of effective planning;
- focuses on how students learn;
- is central to classroom practice;
- is a key professional skill;
- is sensitive and constructive;
- fosters motivation;
- promotes understanding of goals and criteria;
- helps learners know how to improve;
- develops the capacity for self-assessment;
- recognises all educational achievement. (ARG, 2002).

The researchers worked with six schools in two LEAs (Oxford and Medway) to focus on practice in specific areas of classroom assessment which had already been identified as needing development:

- teacher questioning;
- feedback through marking;
- peer and self-assessment by students;
- the formative use of summative tests.

As teachers became focused on particular areas, such as the way in which they ask questions to elicit understanding and development, the researchers emphasised that teachers wanted to learn more about models of learning. This is a shift from the traditional teacher role to one of the teacher as a learner. A further change in the teacher role in AfL was that teachers were enabled to facilitate student learning rather than feeling that they had to cover the curriculum at all costs. In the current climate of pressure to do better, teachers have tended to assume increasing responsibility for their students' learning. (Black et al, 2003).

The research reflects considerable evidence of the positive impact of AfL on summative results; a quarter to a half GCSE grade per student improvement is cited, proof that teachers '*do not have to choose between teaching well and getting good results*'. However, the study does reflect tension between AfL and summative assessment which '*dulled the message about the need to improve, replacing it with information about successes and failures*'. (Black et al, 2003). The question remains about how such a tension can be resolved. A further question also remains about how AfL can be manageable for teachers as part of the future assessment framework.

Assessment by teachers for summative purposes

Wynne Harlen's Eppi-supported review (2004) focuses on the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment, the kind of conditions that affect that process. In a presentation at a GTC assessment seminar prior to the publication of the study, Harlen defined 'reliability' as to '*how accurate the assessment is…If repeated how far would the second result agree with the first*' and defines 'validity' as '*How well what is assessed matches what it is intended to assess*'. (Harlen, Presentation at GTC Assessment Seminar, February, 2004)

The final GTC paper will consider the Eppi study in more detail as part of a consideration about the future balance needed between internal and external assessment. However preliminary findings reflect a degree of unreliability related to TA role. These included:

- the clearer teachers are about the goals of students' work, the more consistently they applied assessment criteria;
- teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching;
- teachers who have participated in developing criteria are able to use them reliably in rating students' work;
- there were considerable differences among teachers in their approaches to TA, variation in the level of TA, and in the difference between TA and standard tests related to the school;
- wide reporting of bias in TA relating to student characteristics, including behaviour (young children), gender, SEN, overall academic achievement and verbal ability that may influence judgement when assessing specific skills. (GTC presentation, 2004).

However, Harlen's study begins to articulate ways in which such unreliability could be addressed. For example, she felt that training:

- should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of the language used;
- should focus on the sources of potential bias that has been revealed by the research.

Harlen also advocated moderation through collaboration, protected time for teachers to meet and the need to develop an 'assessment community' within schools '*allied to increased confidence in the professional judgement of teachers.* (GTC presentation, February, 2004)

These recommendations are particularly significant alongside the assessment proposals being developed in the context of the Interim Report of the 14-19 Reform Group led by Mike Tomlinson. These include the development of Chartered Examiner Status and an Institute of Examiners '*to produce a cadre of skilled and professionally accredited assessors working in schools, colleges and training providers*'. (DfES, 2004). These issues will again be re-visited in the final discussion paper.

Conclusion

This discussion on the role of teachers in the process of assessing pupils leaves many issues unresolved and many questions unanswered. The major one is how a future assessment system can better support teaching and learning while still providing robust evidence of individual and collective pupil progress and performance.

Much of the research included in the discussion reflects the need for teachers to play an enhanced role in assessment across the board. It advocates that teachers should be involved in training and development opportunities to create more collective understanding in schools of assessment and its moderation processes, resulting in the development of embedded assessment cultures.

The final paper in this series will make recommendations by the GTC concerning the principles needed for a future system with a clearer balance between internal and external assessment and for an enhanced role for teacher professional judgement.

References

Assessment Reform Group, 2002, *Assessment for Learning 10 Principles: research-based principles to guide classroom practice*, ARG supported by Nuffield Foundation.

Black, P, Harrison, C, Lee, C, Marshall, B, Wiliam, D, 2003, *Assessment for Learning: Putting it into Practice*, Open University.

Broadfoot, P, 1999, *Empowerment or Perfomativity? English Assessment Policy in the late twentieth century*, BERA annual Conference Paper, Autumn 1999.

Clarke, S, 1996, *The Impact of National Curriculum Statutory Testing at Key Stage 1 and 2 on teaching and learning and the curriculum*, British Journal of Curriculum and Assessment, 1996, Vol 7.

Clarke, S and Gipps, C, 2000, *The Role of Teachers in Teacher Assessment in England 1996-8*, Evaluation and Research in Education, Vol 14, No1, 2000.

Hall, K and Harding, A, 1999, *Teacher Assessment of Seven Year olds in England: A Study of its Summative Function*, Early Years, Vol 20, No 1, Autumn 1999.

Hall, K and Harding, A, 2002, *Level Descriptions and Teacher Assessment in England: towards a community of assessment practice*, Education Research, Vol 44, No 1, Spring 2002.

Harlen, W, 2004, *The Role of the Teacher in Pupil Assessment*, Presentation at GTC Assessment Seminar, February 2004.

Singh, B, 1999, *Formative Assessment: which way now?* BERA Annual Conference Paper, September 1999.

Working Group on 14-19 Reform, 2004, 14-19 *Curriculum and Qualifications Reform*, Interim Report, DfES.

Can assessment by teachers be a dependable option for summative purposes?

Wynne Harlen

Background

Research on assessment has provided insights into how it can affect learning, both positively and negatively. On the positive side, the research review conducted by Black and Wiliam (1988) revealed that assessment, when used formatively, can support learning and raise standards of achievement, particularly of low-achieving students. On the negative side, the review of research on summative assessment by Harten and Deakin Crick (2002), showed that summative assessment that involves high stakes testing reduces students' enjoyment of and motivation for learning in addition to focusing the curriculum and methods of teaching on passing the tests. The overwhelming evidence of impact on motivation, especially the lowering of self-esteem of lower-achieving students, and orienting all students to performance goals rather than learning goals, is particularly serious in the context of the need for developing lifelong learners.

There is also an argument that testing raises standards of achievement. This is based on the increase in scores which often accompanies the introduction of high-stakes testing. Much of this increase, however, can be attributed more to teachers and students becoming familiar with the test requirements than to real improvements in the quality of students' learning. Linn (2000), for example, has shown how changes in the tests are accompanied by a sudden fall in achievement, followed by a rise as teachers begin teaching to the new test. Nevertheless summative assessment is necessary and serves important purposes of providing information as well as summarising students' achievement and progress for their teachers, parents, the students themselves and others who need this information.

We need summative assessment that serves its purposes effectively and without distorting teaching methods and the curriculum. It should reflect the learning outcomes that are important aims in the 'information age' - in particular, learning to learn and motivation for continued learning throughout life (eg OECD, 1999, 2001) and other educational goals that are not readily amenable to formal testing. It also needs to take a form that benefits all students. The use of the information that teachers can gather through their constant contact with students has the potential to meet these needs. As part of their regular work teachers can build up a picture of students' attainments across the full range of activities and goals. This gives a broader and fuller account than can be obtained through any test using a necessarily restricted range of items, and so can be described as a more valid means of assessing outcomes of education (Crooks, 1988; Wood, 1991). Further, in this process the teacher has the opportunity to use this accumulating information to help learning.

So the question arises as to how it may be possible to make more use of teachers' assessment - and less use of tests - for summative purposes, where judgements need to be dependable. Militating against this are widespread assumptions that teachers' assessments are unreliable and subject to bias - despite their use in some countries as a main features of national and state systems. It was to investigate the evidence relating to these assumptions that the review of research, which is the focus of this paper, was carried out.

The questions investigated were:

- what is the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment? .
- what conditions affect the reliability and validity of teachers' summative assessment?

Before summarising the findings it is important to discuss the meaning of the key terms involved.

Matters of definition

The term summative assessment is used to refer to assessment carried out for the purpose of providing a record of a student's overall achievement in a specific area of learning at a certain time. It is the purpose that distinguishes it from assessment described as formative, diagnostic or evaluative (DES, 1987), since some of the methods used for gathering information, such as observation, could be the same for all purposes.

Teachers inevitably have a role in any assessment, but the term assessment by teachers (TA) is used for assessment where the professional judgement of teachers has a significant role in drawing inferences and making judgements of evidence as well as in gathering it. In the present context the term is reserved for assessment by teachers of their own students, thus does not include the role of teachers in setting or marking examination papers. Nor does it refer to school-based assessment where teachers have a role only in gathering evidence that is then marked or graded by others.

The reliability of the result of an assessment, which may be in the form of a test score or summary grade, mark or level, is the extent to which it can be said to be accurate and not influenced by, for instance, the particular occasion or who does the collecting and marking or grading. Thus reliability is often identified as, and measured by, the extent to which, 'if the assessment were to be repeated, the second result would agree with the first'. (Harlen, 2000, p111). So reliability refers to how well the assessment is made, whilst validity refers to what is assessed and how well this corresponds with the behaviour or construct that it is intended to assess. But validity is not a simple concept and various forms of it are identified according to the basis of the judgement of validity. These are identified as 'content', 'concurrent', 'construct', 'consequential'. 'technical' validity, to name only the most common. In searching for an over-arching concept of validity to bring these together there is some agreement (James, 1998) that construct validity subsumes the other types. The argument is that an assessment cannot require the use of the knowledge and skills or other constructs that are supposedly assessed unless there is a clear definition of the domain being assessed, and evidence that in the assessment process the intended skills and knowledge are used by the learners.

The concepts of reliability and validity are not independent of each other in practice.

The relationship is usually expressed in a way that makes reliability the prior requirement. The argument goes that an assessment that does not have high reliability cannot have high validity, for if there is uncertainty about the accuracy of the assessment, and it is influenced by a number of different factors, then the extent to which it measures what it is intended to measure must also be uncertain. Accepting this argument leads to attempts to increase reliability, often by closer and narrower specification of tasks so that the responses to them can be marked with minimum error. This process, however, leads to using a restricted range of evidence and a reduction in validity, since the range of skills and knowledge assessed is not a reflection of what ought to be assessed. On the other hand, if validity is increased by extending the range of the assessment to include outcomes such as higher-level thinking skills, then reliability is likely to fall, since these aspects of attainment are not easily assessed. Clearly, for summative purposes, both reliability and validity have to be considered together, for to attempt to make either as high as possible would lead to a lowering of the other. The concept of dependability, combining validity and reliability, is useful here. Thus in the case of teachers' assessment for summative purposes, where the reason for adopting this approach rather than using tests is to protect construct validity, it is important to consider what is the highest optimum reliability that can be reached whilst preserving construct validity for the products of the assessment to serve its purpose. This would identify the approach giving the most dependable assessment.

The review process

The review of research evidence on the reliability and validity of TA used for summative purposes was a systematic review conducted using the procedures of the EPPI-Centre. These procedures ensure that as wide a range of evidence as possible is identified and that, by applying strict criteria of relevance and quality of research, any conclusions are soundly-based.

The search for studies involved hand-searching journals in the library and on-line, searching electronic databases and using citations and personal contacts. Of a total of 431 studies initially found, 30 were eventually selected for in-depth review. All were written in the English language and 15 were conducted in England, 12 in the United States and one each in Australia, Greece and Israel. All studies were concerned with students between the ages of four and 18. 11 involved primary school students (aged 10 or below) only, 13 involved secondary students (aged 11 or above) only and six were concerned with both primary and secondary students. A summary and the full report, describing the procedures in detail, can be found at the website given in the references for Harlen, 2004.

Main findings of the review

Teachers' assessment in the context of National Curriculum Assessment (NCA) in England and Wales - several studies reported research on the TA introduced as part of the NCA in England and Wales in the early 1990s. Studies of the NCA for students aged six and seven gave evidence of considerable error and of bias in relation to different groups of students. (Shorrocks et al, 1993; Thomas et al, 1998). However as some of this evidence was based on correlations between TA and standards tests or tasks, the interpretation of these data for seven-year-olds, should take into account the variability in the administration of the standard tasks reported by Abbott et al (1994) among others.

Study of the NCA for 11 year olds in England and Wales in the later 1990s shows that results of TA and standard tasks agree to an extent consistent with the recognition that they assess similar but not identical achievements (Reeves et al, 2001). This is despite evidence of variation of practice among teachers in their approaches to TA, type of information used and application of national criteria (Gipps et al, 1996; Hall et al. 1997; Radnor, 1995). The assessment by teachers in the National Curriculum allows evidence to be used from regular classroom work. The evidence is that how teachers go about this varies but this does not in itself necessarily affect the reliability. Teachers vary in their teaching approaches and any less variation in assessment practice would not be expected. Certainly variation according to the nature of the subject and how it is taught, as noted by Radnor (1996), is to be expected if assessment is truly embedded in regular work.

Other evidence indicates that the introduction of teachers' assessment as part of the national curriculum assessment initially had a beneficial effect on teachers' planning and was integrated into teaching (Hall et al, 1997). Evidence collected subsequently, however, suggests that in the later 1990s there was a decline in earlier collaboration among teachers and sharing interpretations of criteria, as support for TA declined and the focus changed to other initiatives (Hall and Harding, 2002).

Portfolio studies in the USA – a number of studies were concerned with the portfolio assessment introduced in some states of the USA during the 1990s. The evaluations conducted by Koretz et al (1991), Koretz et al (1994), Koretz (1998) and by Shapley and Bush (1999) indicated low reliability of these assessments. In these cases, neither the tasks to be included in the portfolios nor the assessment criteria were closely specified; teachers were interpreting generic criteria differently in relation to the specific pieces of work. There is also tentative evidence that estimates of the construct validity of portfolio assessment, derived from evidence of correlations of portfolios and tests, were low (Koretz et al, 1994; Shapley and Bush, 1999). Against this is the evidence that high validity was reported for teachers' judgements guided by check-lists and other materials in a work-sampling system used with young students up to grade 3 (Meisels et al, 2001).

Australian experience of using progressive assessment criteria - evidence from evaluation studies of the use of 'subject profiles' in schools in Victoria, Australia, indicates that finer specification of criteria, describing progressive levels of competency, is capable of supporting reliable TA whilst allowing evidence to be used from the full range of classroom work (Rowe and Hill, 1996). In using subject profiles, teachers rated students' levels of performance in relation to indicators of a number of bands of achievement, in developmental sequence, for each strand of each subject of the curriculum. The results indicated that teachers can make reliable judgements using these indicators. From other sources there is some conflicting evidence as to the relationship between teachers' ratings of students' achievement and standardised test score of the same achievement when the ratings are not based on specific criteria (Hopkins et al, 1985; Sharpley and Edgar, 1986).

Evidence from TA use in specific subject areas - when rating students' oral proficiency in a foreign language, teachers' judgements were found to be consistently more lenient than moderators', but the TA placed students in the same rank order as did experienced examiners. (Good, 1988; Levine et al, 1987).

Evidence from the USA suggests that teachers are able to score hands-on science investigations and projects with high reliability using detailed scoring criteria (Frederiksen and White, 2004; Shavelson et al 1992). UK evidence indicates that teachers' assessment of practica1 skills in science makes a valid contribution to assessment at 'A' level within each science subject but there is little evidence of generalisability of skills across subjects. (Brown et al, 1996).

The research of Hargreaves et al (1996) into teachers' assessment of primary students in the arts shows that the clearer teachers are about the goals of students' work, the more consistently they apply assessment criteria. This supports the finding of Coladarci (1986) that teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching. Both Hargreaves et al (1996) and Frederiksen and White (2004) reported evidence that teachers who have participated in developing criteria are able to use them reliably in rating students' work.

Bias in teachers' assessments - many studies reported that bias in TA relating to student characteristics, including behaviour (for young children), gender, special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills (Bennett et al, 1993; Reeves et al, 2001; Thomas et al, 1998; Shorrocks et al, 1993; Brown et al, 1996, 1998; Delap, 1994, 1995; Wilson and Wright, 1993; Levine et al, 1987). Some variation across schools in the level of TA and in the difference between TA and standard tests or tasks was reported. The evidence was conflicting as to whether this was increasing or decreasing over time. (Reeves et al, 2001; Thomas et al, 1998; Gipps et al, 1996; Hall et al, 1997; Hall and Harding, 2002).

Evidence in relation to the reliability and validity of TA in different subjects was mixed. Differences between subjects in how TA compares with standard tasks or examinations results have been found, but there was no consistent pattern suggesting that assessment in one subject is more or less reliable than in another. (Reeves et al, 2001; Shorrocks et al, 1993; Radnor, 1995; Delap, 1994, 1995; Levine et al, 1987).

It should be noted that much of the evidence of bias in teachers' assessment comes mainly from studies where TA is compared with another measure and based on the questionable assumption that the benchmark measure is unbiased and is measuring the same thing as the teachers' assessment. So, whilst it has been reported that teachers under-rate boys more than girls in mathematics and science as compared with their performance in tests (Reeves et al, 2001), the conclusion might equally be that boys perform above expectation on mathematics and science tests. This could be, for instance, because of boys having better test-taking skills in these areas. Similarly several studies report teachers' assessments of students with special educational needs (SEN) being below their score levels on tests of the same achievements (Reeves et al, 2001; Shorrocks et al, 1993; Thomas et al, 1998). On the assumption that standard tasks are unbiased, there is evidence that TA varies systematically, but the same differences found in relation to gender, first language and SEN have been found by the same authors in the results of standards tasks.

Some key issues

The picture provided by the research is not a clear one. In the practices studied there was evidence of bias and in many cases of low dependability. At the same time, the studies throw light on how the dependability of TA can be improved and how bias can be addressed.

The degree of specification of tasks and assessment criteria - one of the key issues in terms of procedures relates to the extent to which task and assessment criteria are specified. In tests it is taken for granted that marking schemes (or protocols) will match specific items. However, when assessment is not based on specific tasks or items, and instead a variety of tasks may provide the content in which the knowledge or skills to be assessed are shown, the relationship of the criteria to the evidence is more problematic and needs to be made explicit.

Assessment criteria can have a dual function in assessment where evidence is, or can be, taken from a range of activities. One function is to focus attention on relevant evidence; the other is as a basis for interpreting and making judgements of the evidence in terms of the extent to which the criteria are met. For valid assessment it would seem obvious that it is important for the criteria to match the learning goals. This is only the same as matching the assessment tasks, if the tasks are themselves an adequate sample of the full range of goals.

In the Vermont portfolio programme, as Koretz et al (1994) reported, the solution adopted to the problem of low reliability was to have scoring carried out by teachers other than the students' own and who were trained in applying the criteria given that training would be more efficient with scorers gathered together. This takes the Vermont portfolio programme outside the definition of TA used in this review. The alternative, advocated by Shapley and Bush (1999), was to prescribe more closely the work samples so that matching criteria could be used. They suggested that a 'core' set of tasks should be included, thus moving the approach more towards a test than a sample of regular work. It carries the risk of attention being focused on these pieces of work, just as it can be on what is tested by external tests, especially when the outcome is used for a purpose that has high stakes for the teachers.

A compromise is suggested by the evidence from studies in the visual arts, music and science projects that teachers can use criteria consistently when these are designed for specific types of performance and that the more thoroughly teachers understand the criteria the more consistently they apply them.

This indicates the nature of training for teachers that will improve the dependability of their assessments. .

Addressing bias - to be dependable, the sample of behaviour assessed must provide adequate evidence to support the interpretations and judgements based on it. All assessment is subject to error, which can be random or systematic. Random errors in assessment by teachers can have several causes relating to the identification of evidence, the understanding of criteria and the application of criteria. The evidence from the studies reviewed suggests that teachers are more reliable in their assessment when they have a good grasp of the criteria, which will help in identifying relevant evidence as well as in making judgements of it.

Bias is a non-random source of error. In tests this can arise on account of the form in which questions are put and the form in which answers are required. For example gender differences have been reported in relation to open-ended and multiple-choice item forms and in relation to the contextualisation of presented problems. (Gipps and Murphy, 1994; Murphy, 1988; 1993; Parker and Rennie, 1998). In assessment by teachers there may be less bias due to unfamiliar situations, particularly if the assessment is embedded in regular work, but there is more opportunity for knowledge of non-relevant factors, such as behaviour, as well as gender and general performance, unconsciously to influence teachers' judgements.

For assessment by teachers, given that the range of regular activities provides equal opportunities for all to use and develop their knowledge and skills, ideally bias should be eliminated at the point of applying criteria. Efforts to do this include training in careful application of criteria in identifying valid evidence and in making judgements (Gipps and Murphy, 1994).
Bias which exists after judgements have been made can be detected and controlled by moderation and adjustment of the judgements. This can be through comparison with judgements of the same evidence of others, particularly those who have been trained to avoid bias and have experience of looking across a number of teachers' judgements.

Implications for policy and practice

Solutions to the problems of inconsistency in the type of evidence used and in the application of criteria suggested by the studies focused on five types of action. relating to: the specification of the tasks; the specification of the criteria; training; moderation; and the development of an 'assessment community' within the school allied to increased confidence in the professional judgement of teachers.

Some of the implications for assessment policy are:

- when deciding the method, or combination of methods of assessment for summative assessment, the short-comings (such as low reliability and validity, high cost and negative impact) of external examinations and national tests need to be borne in mind;
- the essential and important differences between TA and tests should be recognised by ceasing to judge TA in terms of how well it agrees with test scores;
- there is a need for resources to be put into identifying detailed criteria that are linked to learning goals, not specially-devised assessment tasks. This will support teachers' understanding of the learning goals and may make it possible to equate the assessment tasks with the curriculum;
- it is important to provide for teachers' professional development that addresses the known shortcomings of TA;
- the process of moderation should be seen as an important means of developing teachers' understanding of learning goals and related assessment criteria.

Some of the implications for practice are:

- teachers should not judge the accuracy of their assessments by how far they correspond with test results but by how far they reflect the learning goals;
- there should be wider recognition that clarity about learning goals is a needed for dependable assessment by teachers;
- teachers should be made aware of the sources of bias in their assessments, including the 'halo' effect, and school assessment procedures should include steps that guard against such unfairness;

- schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers is sustained, for example by protecting time for planning assessment, in-school moderation, etc;
- schools should develop an 'assessment culture' in which assessment is discussed constructively and positively and not seen as a necessary chore.

References

Abbott, D., Broadfoot, P., Croll, P., Osborn, M. and Pollard, A. (1994) Some sink, some float: national curriculum assessment and accountability, British Educational Research Journal, 20, 155 - 174.

Bennett, R. E., Gottesman, R. L., Rock, D., A. and Cerullo, F. (1993) Influence of behaviour perceptions and gender on teachers' judgments of students' academic skill, Journal of Educational Psychology. 85,347-356.

Black, P. and Wiliam, D (1998) Assessment and Classroom Learning. Assessment in Education 5(1), 7 -71.

Brown, C. R., Moor, J. L., Silkstone, B. E. and Botton, C. (1996) The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations, Assessment in Education, 3, 377 - 391.

Coladarci, T. (1986) Accuracy of teachers' judgments of students' responses to standardized test items, Journal of Educational Psychology, 78, 141 - 146.

Crooks, T.J. (1988) The impact of classroom evaluation practices on students, Review of Educational Research, 58, 438-481.

Delap, M. R. (1995) Teachers' estimates of candidates' performance in public examinations, Assessment in Education, 2, 75-92.

Delap, M. R. (1994) An investigation into the accuracy of A-level predicted grades, Educational Research, 26, 135-149.

DES (1987). Task Group on Assessment and Testing (TGAT): A Report. London:

Department of Education and Science and Welsh Office.

Frederiksen, J. and White, B. (2004), Designing assessment for instruction and accountability: an application of validity theory to assessing scientific inquiry.

In (Bd) Wilson, M. Towards Coherence between Classroom Assessment and Accountability, 103rd Yearbook of the National Society for the Study of Education part 11. Chicago: National Society for the Study of Education. In press.

Gipps, C., McCal1um, B. and Brown, M. (1996) Models of teacher assessment among primary school teachers in England, The Curriculum Journal, 7, 167 - 183.

Gipps, C. and Murphy, P. M. (1994) A Fair Test? Buckingham; Open University Press.

Good, F. J. (1988) Differences in marks awarded as a result of moderation: some findings from a teachers assessed oral examination in French, Educational Review, 40, 319 - 331.

Hall, K. and Harding, A. (2002) Level descriptions and teacher assessment in England': towards a community of assessment practice, Educational Research, 44.

Hall, K., Webber, B., Varley, S., Young, V. and Donnan, P. (1997) A study of teacher assessment at Key Stage 1, Cambridge Journal of Education, 27, 107 -122.

Hargreaves, D. J., Galton, M. J. and Robinson, S. (1996) Teachers' assessments of primary children's classroom work in the creative arts, Educational Research, 38, 199 - 211.

Harlen, W. (2004) A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes (EPPI-Centre Review). In Research Evidence in Education Library. Issue 3 London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/review_groups/assessment/review three.htm

Harlen, W (2000) Teaching, Learning and Assessing Science 5 - 12. London: Paul Chapman

Harlen, W. and Deakin Crick, R (2002) A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning (EPPI-Centre Review). In Research Evidence in Education Library. Issue 1 London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Hopkins, K. D., George, C. A., and Williams, DD. (1985) The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria, Journal of Educational Measurement, 22, 177-82.

James, M.(1998) Using Assessment for School Improvement Oxford: Heinemann Educational.

Koretz, D., Klein, S, and Shepard, L.A. (1991) The Effects of High-Stakes Testing on Achievement: Preliminary Findings about Generalization across Tests. Paper presented at the Annual Meetings of the American Educational Research Association (Chicago, IL, April 3-7, 1991) and the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).

Koretz, D (1998) Large-scale portfolio assessment in the US: evidence pertaining to the quality of measurement. Assessment in Education 5 (3), 309 - 334

Koretz, D., Stecher, B. M., Klein, S. P. and McCaffrey, D. (1994) The Vermont Portfolio Assessment Program: findings and implications, Educational Measurement: Issues & Practice, 13,5- 16.

Levine, M. G., Haus, G. J and Cort, D. (1987) The accuracy of teacher judgment of the oral proficiency of high-school foreign language students, Foreign Language Annals, 20, 45-50.

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

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y. and Atkins-Bumett, S. (2001) Trusting teachers' judgments: a validity study of a curriculum-embedded performance assessment in kindergarten to Grade 3, American Educational Research Journal, 38, 73-95.

Murphy, P. M. (1988) Gender and assessment. Curriculum, 9, 165 -171.

OECD (2001) Knowledge and skills for life: first results from the PISA 2000. Paris: OECD

OECD (1999) Measuring Student Knowledge and Skills. Paris: OECD

Parker, L. H and Rennie, L. J. (1998) Equitable assessment strategies. In B. J. Fraser and K. G. Tobin (eds). International Handbook of Science Education. 897910.

Radnor, H. A. (1996) Evaluation of Key Stage 3 Assessment Arrangments for 1995. Final Report University of Exeter, Exeter, pp. 181.

Reeves, D. J., Boyle, W. F. and Christie, T. (2001) The relationship between teacher assessment and pupil attainments in standard test/tasks at key stage 2, 1996 8, British Educational Research Journal, 27, 141-160.

Rowe, K. J. and Hill, P. W. (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles', Assessment in Education, 3, 309-352.

Shapley, K. S. and Bush, M. J. (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience, Applied Measurement in Education, 12, 11-32.

Sharpley, C. F. and Edgar, E. (1986) teachers' ratings vs standardized tests: an empirical investigation of agreement between two indices of achievement, Psychology in the Schools, 23, 106 - 111.

Shavelson, R. J., Baxter, G. P. and Pine, J. (1992) Performance assessments: political rhetoric and measurement reality, Educational Researcher, 21, 22 - 27.

Shorrocks, D. Daniels, S., Stainton, R. and Ring, K (1993) Testing and Assessing 6 and 7 year olds. The evaluation of the 1992 Key Stage 1 National Curriculum Assessment. National Union of Teachers and Leeds University School of Education.

Thomas, S., Madaus, G. F., Raczek, A. E. and Smees, R (1998) Comparing teacher assessment and the standard task results in England: the relationship between pupil characteristics and attainment, Assessment in Education. 5,213 - 246.

WiIson, J. and Wright, C. R. (1993) The predictive validity of student self-evaluations, teachers' assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachia schools, Educational & Psychological Measurement, 53,259-70.

Wood, R. (1991) Assessment and Testing: a Survey of Research. Cambridge: University Press.

Fairness in assessment

Caroline Gipps and Gordon Stobart

Introduction

This paper addresses some of the key issues in fair assessment: equal opportunities, bias and validity. We focus particularly on tests, though we also apply our argument to teacher assessment. Equal opportunities in assessment relate to two issues; what we commonly call 'bias' in the test itself, and fairness in the comparison: have the groups of pupils being tested had the same opportunities to learn? Fairness and equity are used as interchangeable terms, with equal opportunities as one component of what constitutes equity. The key question is: can we create an assessment system that is fair to all learners? The answer is: no - but we can make it fairer. We do this by being clear about what we are assessing, by identifying and dealing with possible sources of bias and unfairness.

Most research on equity issues in assessment has focused on tests and examinations; there has been little work done on equity issues in teacher assessment. For example, we know that teacher expectation can affect the curriculum and learning experiences offered to children. There is clear evidence that teachers offer a different curriculum to children for whom they hold low and high expectations (Tizard et al 1988; Troman 1988, Harlen 2004). While high teacher expectation is good and can enhance pupil performance, the opposite also holds true. So, one question is, can teacher expectation have an effect on teacher assessment?

Equity and testing

It is important to remember that external testing has historically been seen as an instrument of equity. '*Examinations were the obvious method of attacking patronage, the hitherto dominant mode of recruitment to all forms of government*' (Sutherland, 1996, p.16). The notion of the standardised test as a way of offering impartial assessment is a powerful one, though if equality of educational opportunity does not precede the test, then the 'fairness' of this approach is called into question. So these 'fair' 19th century selection examinations invariably excluded women from taking them.

Bias is a term widely used in relation to assessment and is generally taken to mean that the assessment is unfair to one particular group or another. This rather simple definition, however, belies the complexity of the underlying situation. Differential performance on a test, ie where different groups get different score levels, may not be the result of bias in the test; it may be due to real differences in performance among groups which may in turn be due to differing access to learning, or it may be due to real differences in the group's attainment in the topic under consideration. It is also possible to have unequal group outcomes that may be seen as fair. An example would be where there are group differences in the *application* to learning and preparation, where each had similar resources and *opportunities.* The philosopher John Wilson has argued '*Education is not (only) something that can simply be given to people and distributed equally or unequally, like cake. To be educated is not just to have received something but also to have done something... there is always what we may call the question of uptake: whether the individual makes use of whatever opportunities or resources he may be given*' (1991, p. 223).

The question of whether a test is biased or whether the group in question has a different underlying level of attainment is clearly extremely difficult to answer. Wood (1987) describes these different factors as the opportunity to acquire talent (access issues) and the opportunity to show talent to good effect (fairness in the assessment). In the USA tests have been seen to be denying opportunities for advancement, particularly for black students. In the post-1965 Civil Rights legislation era, critics of 'advancement through testing' were pointing out that opportunities to acquire talent, or to be able to show it to sufficient effect in tests and examinations, were not equally distributed (Wood, 1987; Orfield and Kornhaber, 2001). In other words, these tests were biased in favour of the dominant social group.

The traditional psychometric view has been that technical solutions can be found to solve problems of equity with the emphasis on using elaborate techniques to eliminate biased items (Murphy, 1990; Goldstein, 1993). A limitation of this approach is that it does not look at the way in which the subject matter is defined (ie the constructs around which test items are designed); nor at the initial choice of items from the thus-defined pool; nor does it question what counts as achievement. It simply 'tinkers' with an established selection of items. Focusing on bias in tests, and statistical techniques for eliminating 'biased' items, not only may confound the construct being assessed, but has distracted attention from wider equity issues such as actual equality of access to learning, 'biased' curriculum, and inhibiting classroom practices.

Fairness

Most tests and examinations are amenable to 'coaching' and pupils who have very different school experiences are not equally prepared to compete in the same test situation. Furthermore, pupils do not come to school with identical experiences and they do not have identical experiences at school. We cannot, therefore, expect assessment to have the same meaning for all pupils. However, the stakes and purpose of the assessment are relevant here as Linn et al (1991) argue: "*On a non-threatening assessment... it is reasonable to include calculator-active problems even though student access to calculators may be quite inequitable. On the other hand, equitable access would be an important consideration in a calculator-active assessment used to hold students or teachers accountable*" (1991, p. 17). What is important is to have a fair approach where the concerns, contexts and approaches of one group do not dominate. This, however, is by no means a simple task; for example, test developers may be told that they should avoid any context which may be more familiar to males than females or to the dominant culture. But there are problems inherent in trying to *remove* context effects by doing away with passages that advantage males or females, because it reduces the amount of assessment material available. De-contextualised assessment is anyway not possible, and complex higher-order skills require drawing on complex domain knowledge.

For design of tests in a multicultural society Shohamy (2000) has proposed three models of how the contributions of different groups are treated;

- *The assimilative model.* In this there is no appreciation of an immigrant's (sic) previous knowledge; the task is to master the new knowledge associated with the dominant group. There may be recognition that this takes time to acquire and allowances may be made to ease the process ('pain-killers');
- *The recognition model.* In this there is recognition and appreciation of the different knowledge and viewing of it as valuable - a situation in which groups are credited for this knowledge and encouraged to maintain it;
- *The interactive model.* In this the knowledge of the 'different' groups affects and influences the dominant group and thus enriches existing knowledge.

While we might aspire to the interactive model, Shohamy (representing the highly diverse Israeli culture) is not optimistic:

"*Even in societies that recognise multiculturalism as part of society there is rarely recognition of the specific and unique knowledge of the different groups in schools... educational leaders continue to strive for homogenous knowledge to be owned by all. This is even more apparent in educational assessment. In a number of situations there is a gap between curricula and assessment as curricula may, at times, contain statements and intentions for the recognition of diverse knowledge, yet the tests are based on homogeneous knowledge.*" (p. 3)

Perhaps one litmus test of where an assessment system is in relation to these models is in the attitude to language: how much linguistic diversity does the assessment system reflect? For example, should we:

- assess in only the main language of the culture (eg. England);
- offer the same tests/qualifications in two or more languages (eg. Wales).

Both options bring benefits and costs. In the monolingual approaches an issue is the accessibility of tests for those who are not using their first language, particularly if this is combined with cultural assumptions in their content. Politt et al (2000) provide a case study example of how the monolingual assumptions of mathematics test writers interfered with understanding of an Urdu-speaking student taking a mathematics test in English. In Urdu the number of hours 'in a day' (*din*) is 12 (with day-night, *dinraath*, being 24 hours) and there are two words for 'height' (from the ground; of the object) – with both ambiguities capable of generating 'wrong' answers to everyday 'how long would it take…?' and 'how high is... ?'

Ruddock and Evans (2000) also provide some construct dilemmas in providing English-Welsh translations. These include differences in demand, for example giving the meaning in science of 'hibernation' may be a less demanding item in Welsh, in which it is translated 'sleeps in winter'.

Forms of assessment

We are now well aware that the form of assessment can differentially affect results for different groups. In England there has been far more analysis of this in relation to gender than to ethnicity. We know that during compulsory schooling (up to 16 years) girls are likely to outperform boys on tasks which involve open-ended writing, particularly when this involves personal response. The gap narrows if the responses are fixed-choice or short-answer (Gipps and Murphy, 1994). Even within multiple-choice tests, traditionally seen as favouring boys, there are differential response patterns. Carlton (2000) has shown that in such tests females perform better than males, matched for ability, on questions in which the content is a narrative or is in a humanities field and when the content deals with human relationships. As the context of an item grows larger the relative performance of females also improves. Males outperform females on questions relating to science, technical matters, sports, war or diplomacy. We also know that where examinations have a coursework element the performance of girls is likely to be more consistent, though the effect this has on final grades has often been overstated (Elwood, 1995).

We know less about other aspects of the form of assessment, particularly in relation to ethnicity. For example *oral* assessment plays little part in the examination system in England outside examining languages. Does the emphasis on written response disadvantage groups who place more emphasis on oral communication in their culture? Rudduck (1999) has raised this in relation to Afro-Caribbean boys in England who, as a group, often perform less well than others on written examinations.

An agenda for assessment

So, to return to our definition of equity, how do we ensure that assessment practice and interpretation of results is as fair as possible for all groups? As Gipps and Murphy (1994) and Willingham and Cole (1997) argue, consideration of the way in which a construct is tested is crucial. We need to encourage clear articulation of the test developers' constructs on which the assessment is based, so that the construct validity may be examined by test-takers and test-users. The requirement is to select assessment content that *accurately reflects the construct*, even if it produces gender/ethnic group differences, and to avoid content that is *not relevant to the construct* and could affect such differences. The ethics of assessment demand that the constructs and assessment criteria are made available to pupils and teachers and, in any case, this is consonant with enhancing construct validity. We also need to define the context of an assessment task as well as the underlying constructs to make sure they reflect what is taught. The involvement of those with a 'minority' background is crucial here.

Baker and O'Neil (1994) report some uncomfortable findings on performance assessment (ie assessment involving tasks and projects) in relation to ethnic minorities: "*The minority community's perception of the self-evident merit of performance assessment deserves additional exploration… the major assertion was that performance-based assessment reform is a creation of the majority community intended to hold back the progress of disadvantaged children. Performance-based assessment is obviously grounded in a different instructional model, one in which the majority of teachers of disadvantaged children may be unprepared... Although most of the concerns were articulated by African-Americans, there was also the early recognition that much of performance-based assessment required strong language skills by students to explain or document their accomplishments. It is undeniable that in the US, performance assessment is a white, middle-class venture, promoted by high-achieving people, disproportionately women. Minority communities must not once again become unwilling recipients of innovations which other believe are good for them.*" (pp 13-14)

An important approach to offering fairness is to use, within any assessment programme, a range of assessment tasks involving a variety of contexts; a range of modes within the assessment; and a range of response format and style. This broadening of approach, though not always possible, is most likely to offer pupils alternative opportunities to demonstrate achievement if they are disadvantaged by any one particular assessment in the programme. (Linn, 1992)

Indeed, this is included in the Criteria for Evaluation of Student Assessment Systems by the USA National Forum on Assessment (NFA):

- to ensure fairness, students should have multiple opportunities to meet standards and should be able to meet them in different ways;
- assessment information should be accompanied by information about access to the curriculum and about opportunities to meet the Standards;
- assessment results should be one part of a system of multiple indicators of the quality of education. (NFA, 1992, p. 32)

If we wish pupils to do well in tests/exams we need to think about assessment which elicits an individual's best performance (after Nuttall, 1987). This involves tasks that are concrete and within the experience of the pupil (an equal access issue) presented clearly (the pupil must understand what is required of her if she is to perform well) relevant to the current concerns of the pupil (to engender motivation and engagement) and in conditions that are not threatening (to reduce stress and enhance performance) (Gipps, 1994). This is where teacher assessment *can* be more equitable since it is under the teacher's control. (Gipps, 1994)

As good assessment practice we should be:

- using assessment that supports learning and reflection, including formative assessment with feedback;
- designing assessment that is open and linked to clear criteria (rather than relying upon competition with others);
- including a range of assessment strategies so that all learners have a chance to perform well.

Using a range of assessment processes, together with clarity and openness about what is being assessed and how, is not only more equitable, but also supports learning. This is as true for teacher assessment as it is for examinations and tests.

References:

Baker, E. and O'Neil, H. F. (1994) Performance, Assessment and Equity: a view from the USA. *Assessment in Education*, Vol. 1, No. 1, pp. 11-26.

Carlton, S. T. (2000) Contextual Factors in Group Differences in Assessment, *paper presented at 26th IAEA Conference*, Jerusalem.

Elwood, J. (1995) Undermining gender stereotypes: examination and coursework performance in the UK at 16. *Assessment in Education*, Vol. 2, No. 3, pp. 282-303.

Gipps, C. (1994) Developments in Educational Assessment: what makes a good test? *Assessment in Education*, Vol. 1, No. 3, pp. 283-291.

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

Goldstein, H. (1993) *Assessing group differences*, Oxford Review of Education, Vol. 19, pp. 141-150.

Linn, M. C. (1992) Gender differences in educational achievement, *Sex Equity In Educational Opportunity, Achievement and Testing*, ETS, Princeton, NJ.

Linn, R. L., Baker, E. and Dunbar, S. (1991) Complex, performance-Based Assessment: Expectations and Validation Criteria, *Educational Researcher,* Vol. 20, No. 8, pp. 15-21.

Murphy, P. (1990) *Gender difference – implications of assessment curriculum planning*, BERA, Roehampton, August.

National Forum on Assessment (NFA) (1992) Criteria for evaluation of student assessment systems, *Educational Measurement: issues and practice*, Spring.

Nuttall, D. (1987) The Validity of Assessments, *European Journal of Psychologyof Education*, Vol. 11, No. 2, pp. 108-118.

Orfield, G. and Kornhaber, M. L. (2001) *Raising Standards of Raising Barriers?* Century Foundation, New York.

Pollitt, A., Marriott, C. and Ahmed, A. (2000) Language, Contextual and Cultural Constraints on Examination Performance. *Paper presented at26th IAEA Conference,* Jerusalem.

Ruddock, G. and Evans, S. W. (2000) Working in Two Languages (or Two Variants of the Same Language), *Paper presented at 26th IAEA Conference*, Jerusalem.

Rudduck, J. (1999) Seminar presentation, QCA, London.

Shohamy, E. (2000) Educational Assessment in a Multi-Cultural Society: Issues and Challenges, *Paper presented at 26th IAEA Conference*, Jerusalem.

Sutherland, G. (1996) Assessment: Some Historical Perspectives in H. Goldstein & T. Lewis (Eds) Assessment: Problems, Developments and Statistical Issues, London, John Wiley & Sons.

Tizard, B. et al. (1988) *Young Children at School in the Inner City*. Lawrence Erlbaum Associates, Hove.

Troman, G. (1988) 'Getting it right: selection and setting in a 9-13 middle school'. *British Journal of Sociology of Education*, Vol. 9, No. 4.

Willingham, W. and Cole, N. (1997) *Gender and Fair Assessment*, Lawrence Erlbaum Associates, NJ/London.

Wilson, J. (1991) Education and Equality: some conceptual questions, *Oxford Review of Education*, Vol. 17, No. 2, pp. 223-230.

Wood, R. (1987) *'Assessment and Equal Opportunities'*. Text of public lecture at ULIE (11 November 1987).

Internal and external assessment: What are we talking about?

Mary James

Internal to what?

When we use the adjectives 'internal' and 'external' we have in mind a bounded entity that has an inside and an outside. The distinction between internal and external assessment assumes this also, yet identifying the entity is not as obvious as it might seem. 'It's the school, stupid!' might be the first response. But if by 'internal assessment' we mean assessment by teachers, then strictly speaking they are internal assessors in their own classrooms with their own students but may be regarded as outsiders if given a role in assessment, say as moderators, of the achievements of students in colleagues' classes. In some circumstances this is unproblematic but in a climate where accountability matters, and performance management is a feature of schools, issues of status, role and differential power need to be recognised and dealt with. Assessment is not merely a technical matter; it is a deeply social practice.

The distinction is similarly muddy when considering the outside boundary of schools. In the past the physical architecture of schools defined their bounds but in the 21st century the term 'school community' is increasingly used. At the very least this includes governors and, especially, parents as well as teachers, students and support staff. What implications does this have for internal assessment? Should parents have a role?

One interpretation of the distinction, therefore, puts an emphasis on role: who is involved in assessment and in what ways. The role of teachers is of special interest. Another interpretation is to focus on internal or external purposes, which may or may not imply a central role for teachers. Certainly, teachers can, and are, involved in making assessments for external purposes, although it is doubtful whether they can be denied a role in assessment for internal purposes. But then we also need to be clear about what we mean by internal and external purposes.

Clarifying purposes

Books on assessment usually have an introductory chapter on assessment purposes because a key criterion of quality in assessments is their 'fitness for purpose'. However, the lists of purposes quoted are often as diverse as their authors. Sometimes they are brief, reduced, following TGAT (1988), to formative, summative and sometimes diagnostic and evaluative. Other times they are much longer including such things as screening, allocating resources, feedback to students, target setting, curriculum planning, student grouping, prediction, guidance, certification and accreditation, monitoring etc. In recognition that these two lists are not quite of the same order, Gipps and Stobart (1993) made a helpful distinction between the 'purposes' and the 'uses' of assessment: the one focusing on the intention behind the assessment process and the second focusing on the actual use made of the results. This is a temporal distinction which enables us to look at purposes in prospect and retrospect. It is also helpful in sorting them into clusters.

Another way of grouping purposes is according to whether they relate to policies that are internal or external to the school (James, 1998: 24). For example, diagnosis, feedback to students and teachers and individual target-setting are purposes internal to the school, whilst certification, selection, monitoring standards, evaluation of school performance and accountability relate to purposes that are largely external to the school. Some, of course, can be both. These two kinds of distinction (one temporal and the other spatial) are amalgamated in the chart below (Table 1).

Table 1

Intention (purpose in process)	**Realisation** (purpose in process)	**Use** (purpose as product)
Assessment for learning	Formative assessment (incorporated into the processes of teaching and learning to improve them; see Black et al. 2002)	Improvement of learning processes and outcomes [internal]
		
Assessment of learning	Summative assessment (summing-up achievements at a given point)	Diagnosis [internal] Tracking [internal] Grouping [internal] Target setting [internal] Reporting [internal] Certification [internal and external] Selection [external] Evaluation [external] Monitoring standards [external]

Note: It is possible for formative assessment to feed into summative assessment and for summative assessments to be used formatively but this is not unproblematic (see Harlen and James, 1997; Wiliam, 2000; Black et al, 2002).

But why bother to classify purposes in these ways? The simple answer is that it helps us evaluate the balance of assessment activities and decide whether the weight of some against others needs adjustment. As can be seen in Table 1, the uses attached to assessment of learning (summative assessment) are far more numerous than for assessment for learning (formative assessment) although the importance of the latter may be greater in the long term, and especially for lifelong learning. What I wrote in 1998 therefore still seems to be the case:

"*In recent years there has been increased external pressure to make schools accountable through the publication of performance tables etc. Thus there has been new emphasis on assessment data collection for monitoring, evaluation, marketing and accountability purposes. Statutory requirements cannot be avoided and there is a natural tendency to give them priority – to put what has to be done first. This can create an imbalance in a school's assessment procedures with internal purposes either sacrificed or made secondary to external purposes. Schools should be watchful of this because their aims for the education of their students are unlikely to be well served if they only pay regard to external demands. Schools' assessment policies require a balance of assessment purposes.*" (James, 1998: 25)

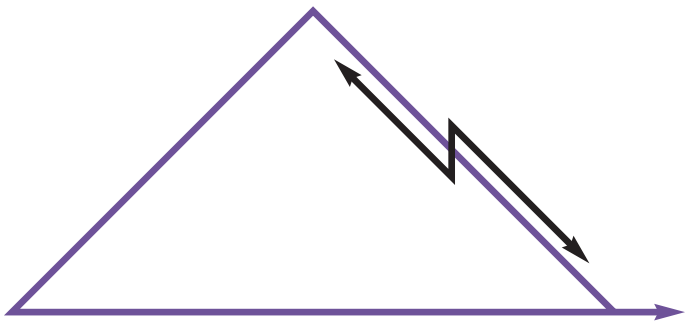
Assessment as an activity system

Another way of evaluating the balance between internal and external assessment practices might be to examine them through the lens of Activity Theory, which derives mostly from European cultural-historical psychology (especially Vygotsky, Leont'ev and Luria) and is developed extensively in organisational settings by Engeström (1999). The attraction of Activity Theory is that enables us to take an activity and see it as a system which is both individual and social and involves process and product. It can be used for description, analysis and prediction in relation to almost any activity in which people use tools, material or conceptual, to achieve some end. Assessment is an activity of this kind. Moreover, it enables us to regard activity systems as multiple, fluid, interlocking, but manageable, rather than as overarching structures in which individual, human agency has little or no power. This is important in relation to assessment because teachers can so easily believe that they are powerless to affect change in assessment systems which they see as imposed from 'above'.

From the perspective of Activity Theory, summative assessment by teachers, to take one example, can be construed as a collective activity system and can be represented in the following way:

Mediating artefacts and conceptual tools:

- assessment tasks
- criteria
- exemplars, mark schemes
- guidelines for moderation



Subjects: teachers

Object: to sum up achievement at a given time

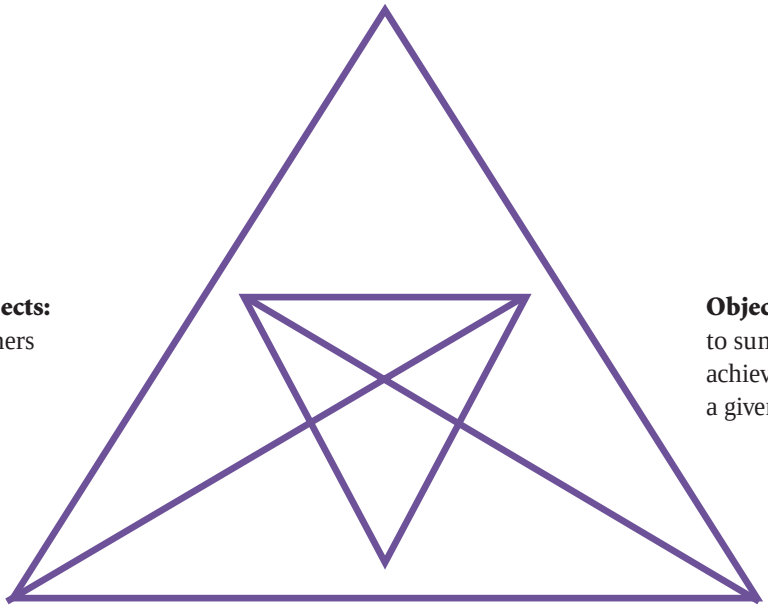
Outcome: grades, scores for particular uses

The subjects of this activity are the teachers because they are carrying out the activity. The object (of the activity, ie what they are working on) is to engage in an assessment process to sum up students' achievement/attainment at a given time. In order to do this, teachers use mediating artefacts in the form of assessment tasks, tests, exemplars or guidance materials, and mediating conceptual tools in the form of criteria. The outcome of this process is the summative judgement in the form of marks, grades or scores which can then be used for purposes of tracking, certification, selection etc. The links represented by the sides of the triangle are all-important, for they indicate either congruence or contradiction (eg if the tools are invalid tests, the object may not be satisfactorily achieved). A contradiction is indicated by the lightning-shaped arrows which become a focus for critical-reflection with a view to change.

This triangle is the classical way of representing activity but it fails to acknowledge the socio-cultural (situated) nature of actions. Activity theory attempts to depict the collective nature of activity by representing another set of components that underpin this triangle. In summative assessment, as in most other spheres, teachers do not work alone: they are part of, and are influenced by – in positive or negative ways – a community in which they work. The activity is also fashioned by rules or norms, eg conventions governing the choice of tasks that are thought to give the most valid and reliable information, or assumptions about the relationship between teachers' assessment processes and test results. Finally, any teacher's work is bounded by the division of labour which is adopted in pursuit of the object, so, who marks tests, and whether students have a role in the assessment activity, all serve to define or to put limits around the teacher's own field of action. The diagram below illustrates this more complex model of summative assessment by teachers as an activity system.

Mediating artefacts and conceptual tools:

- assessment tasks
- criteria
- exemplars, mark schemes
- guidelines for moderation



Subjects:
teachers

Object:
to sum up achievement at a given time

Outcome:
grades, scores for particular uses

Rules
Norms aboutwhat count as valid and reliable tasks, how judgements are made using criteria, moderation, and subject conventions

Community
Other teachers involved in summative assessment at Key Stages or in subjects domains

Division of labour
Distinctions between those who devise tasks, mark and moderate

The value of such a model lies in the extent to which it provokes debate about the elements in an activity system (whether all salient features are taken into account) and the relationships among these features, in particular, whether the system is beset by unresolved tensions, where these are located, and what might be done about them. In other words, its practical usefulness may depend on (a) whether the need to think about each of the components helps obtain a comprehensive view of how objects and outcomes are, or are not, achieved and (b) whether the need to think about each of the links helps achieve insight into how the system actually works, and/or can be changed. With regard to assessment, the concepts of rules, community and division of labour reinforce the idea that distinguishing between internal and external is not a simple matter (see above).

Internal and external elements of assessment activity

If one takes some of the elements of activity systems described above (eg subjects, objects, outcomes, artefacts, rules, communities, division of labour) it is then possible to compare different kinds of assessment (activity systems) and identify some of the inherent tensions. This focus on different aspects of activity makes it clear that it is more valid to describe elements as internal or external, rather than the activity in its entirety. In Table 2, three different but inter-related activity systems are analysed: National Curriculum tests and examinations; National Curriculum summative teacher assessment (TA); formative teacher assessment (ta).

Table 2:

Element of activity	**NC tests and exams**	**NC TA** (summative)	**TA** (formative)	**Questions**
Choice of assessment objectives	External	External (ATs)	Internal	What is the scope for students to choose their own objectives?
Choice of task	External	Internal (but often using externally developed tests)	Internal	What is the role for well-developed tasks in formative assessment?
Administration	Internal	Internal	Internal	Is external test administration a good use of teachers' professional skills and time?
Development of criteria and mark schemes	External	External (level descriptions)	Internal but usually with reference to external	Are criteria the essential link between formative and summative assessment?
Judgement	External	Internal	Internal	Are quantitative judgements (levels, grades, scores) overused?
Validation	External	Combined (through moderation or use of reference tests)	Internal (if at all)	Should involvement in moderation processes be a key focus for professional development?

Is the balance right?

According to the analysis in Table 2, in contrast to that in Table 1, the balance between internal and external elements of current assessment activity systems is more equal. Nevertheless, it is undeniable that summative assessment for accountability purposes (performance management of teachers, evaluation of schools, monitoring the system as a whole) carry the highest 'stakes' and here external activity dominates. This continues to send the message that internal assessment processes cannot be relied upon.

Can the balance be shifted?

Students, parents, receiving teachers, and the public more generally, need to have confidence in teachers' assessment skills and judgements. But, as a team of Canadian evaluators noted (Earl, et al. 2000), 'assessment literacy' is not strong among teachers in England. Neither, according to Ofsted (2003: 3-4), is their practice, especially in terms of internal assessment for learning. Hence, we now have a major thrust in this direction through the revised Key Stage 3 Strategy and the Primary National Strategy, both of which have assessment for learning as a key strand. The Government's recent ideas about Personalised Learning also make assessment for learning a core feature [1]. Whether the strategies for achieving these ends presage more ring binders, consultants etc., and whether these will 'work', remains to be seen.

Certainly, training in assessment (principles and practice) needs to be built into the core of teacher training, both initial teacher education and continuing professional development. What is being asked of teachers is not trivial; it demands not only new knowledge and skills but fundamental changes in the way they think about their role as teachers and the ways in which they go about teaching for learning. In this respect, I take the view that practice will be enhanced if teachers are equipped with understanding and a repertoire of practice (why and how), rather than prescriptions about what to do, for the simple reason that they will need to 'engage' with the ideas rather than simply 'adopt' them. As with students' learning, teachers' learning needs to be mindful. They also need the freedom to make professional choices in the light of their understanding of the different contexts in which they work. Colleagues and I expect that our current ESRC TLRP project 'Learning how to learn – in classrooms, schools and networks' will have something to report on these matters shortly (see http://www.learntolearn.ac.uk).

Teachers also need to be supported by well-developed resources (mediating tools and artefacts). These can include diagnostic instruments or task banks in subject areas and for students of different ages and prior experience. Scotland has pursued this course and another project from the Teaching and Learning Research Programme has produced diagnostic assessment materials for science teaching (see http://www.tlrp.org/pub/research/no1.pdf). Similarly, materials that exemplify what can be achieved by students at different stages and in different contexts will be important for raising expectations of teachers, students and parents. Equally, support for the development of sensitive but robust moderation procedures will be vital. There is an obvious role here for LEAs, consortia, clusters or networks of schools. In Wales it is proposed that secondary schools should be accredited for moderation purposes (DARG, 2004) although the Secondary Heads' Association advocates the accreditation of individual teachers as 'chartered examiners' (SHA, 2003).

However, the priority must surely be the promotion of Assessment for Learning. Without learning, both as process and as outcome, educational assessment – internal or external - serves no purpose.

---

[1] Personalised Learning entered policy discourse when Tony Blair, prime minister, mentioned it in his speech to the 2003 Labour Party conference, **http://politics.guardian.co.uk/labour2003/story/0,13803,1052752,00.html**. The reference is in part two of the speech. Schools minister David Miliband spoke on personalised learning at the North of England education conference in Belfast, January 2004 and in a further speech in May 2004. They can be seen at **http://www.dfes.gov.uk/speeches/search_detail.cfm?ID=95** and **http://www.dfes.gov.uk/speeches/search_detail.cfm?ID=118** respectively.
David Hopkins, former head of standards and effectiveness at DFES, spoke on personalised learning at the ConfEd conference in January 2004. His PowerPoint presentation is at **http://www.virtualstaffcollege.co.uk/download/David%20Hopkins.ppt**

References

Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2002) Working Inside the Black Box. London: King's College London.

DARG (2004) Learning Pathways Through Statutory Assessment: key stages 2 and 3: Daugherty Assessment Review Group: Final Report. Cardiff: Welsh Assembly Government.

Earl, L. et al. (2000) Watching and Learning: OISE/UT Evaluation of the Implementation of the National Literacy and National Numeracy Strategies. Toronto: OISE/UT/

Engeström, Y., Miettinen, R. and Punamäki, R-L. (Eds) (1999) Perspectives on Activity Theory. Cambridge: Cambridge University Press.

Gipps, C. and Stobart, G. (1993) Assessment: A Teachers' Guide to the Issues. London: Hodder and Stoughton (second edition).

Harlen, W. and James, M. (1997) 'Assessment and Learning: differences and relationships between formative and summative assessment', Assessment in Education, 4(3) Autumn, pp. 365-379.

James, M. (1998) Using Assessment for School Improvement. Oxford: Heinemann.

Ofsted (2003) The National Literacy and Numeracy Strategies and the Primary Curriculum. London: Ofsted.

SHA (2003) Examinations and Assessment: proposals by the Secondary Heads Association for a radical reform of examinations and assessment: policy paper 2. Leicester: SHA.

TGAT (1988) National Curriculum: Task Group on Assessment and Testing: A Report. London: DES/WO.

Wiliam, D. (2000) 'Recent development in educational assessment in England: the integration of formative and summative functions of assessment.' Paper presented at the first meeting of the Scientific Advisory Board for the Swedish national Mathematics and Science Tests held at the University of Ume?, May.

Internal and external assessment: What is the right balance for the future?

GTC paper 3: Advice to the Secretary of State

Summary of recommendations:

Action by government

The government should conduct a fundamental review of statutory and non-statutory assessment at Key Stages 1-3 similar to the assessment part of the current 14-19 Reform Group.

The review of Key Stages 1-3 assessment should also consider the most effective ways of making formative and summative assessment information available for a range of uses and audiences. This should include a review of the future role of published performance tables.

The government should continue to invest in assessment for learning through the National Strategies working with the Assessment Reform Group (ARG). It must ensure that LEAs can support schools' and teachers' capacity to conduct assessment effectively, and should aim to revitalise learning in the process.

The government should develop a teacher assessment model that focuses solely on formative assessment to avoid the distortion of the function. This would also enhance pupil learning and help prevent the imposition of unmanageable extra workload on teachers.

Government investment in assessment for learning should include the provision for LEAs to develop local assessment networks and professional knowledge.

The government should consider the proposal for banks of summative activities and tasks to be developed from which individual teachers could select and use with their pupils at a time determined by them. This could offer a new way of providing information on pupil achievement at a particular point which could be summarised for the school, the LEA, parents and the pupils themselves.

The Council believes that the summative assessment of individual pupils is separated from the collection of summative data to be used for national monitoring. Schools should be required to provide summative data for monitoring LEA and national standards involving a rolling programme of samples of pupil cohorts.

Changes to the assessment model should provide opportunities for schools to develop a new accountability relationship with parents which is based on a richer dialogue than external test results.

The New Relationship with Schools (NRwS), the greater emphasis on school self-evaluation, better use of performance data by schools and the development of the School Profile should be promoted as a means for schools to develop a new accountability relationship with pupils and parents.

Action by the General Teaching Council

The Council wishes to work with the Teacher Training Agency (TTA) and other partners to develop teachers' assessment skills as part of the professional standards underpinning Qualified Teacher Status (QTS), induction and continued professional development (CPD).

The Council wishes to work with the DfES and QCA to ensue that specialist expertise in assessment is reflected in teachers' career paths and pathways and with other parts of the professional standards framework.

The Council will work with the DfES, QCA and other partners to develop the e-assessment agenda and to ensure that it best serves teaching and learning.

The Council will propose to the DfES that it works with a small number of schools and teachers within LEA CPD projects to develop its teacher assessment model further, and to encourage the creation of local assessment networks of expertise.

Action by other partners

The QCA and awarding bodies should develop banks of summative activities and tasks from which teachers can select. Chartered assessors and examiners should be involved in this work.

The National Assessment Agency should ensure that the role of the Chartered Assessor/Examiner is pivotal to any consideration of the proposal for accrediting schools to assess summative tests and tasks.

The QCA and awarding bodies should develop online tests as part of the creation of banks of summative assessment tests/activities, which could also be involved in the accreditation of schools to carry out summative tests.

The QCA and other partners should also develop adaptive testing, electronic moderation of test data and new procedures for marking to support further improvements to the administration of summative assessment.

Introduction

This third and final paper is based on evidence presented at the GTC seminar on the future balance needed between internal and external forms of assessment. The paper also contains the Council's agreed recommendations on the agenda for change that the Council believes need to be taken up by the government.

The accountability context

The 1988 and 1992 Acts which introduced a National Curriculum and Assessment Framework, competition between schools to attract pupils and parents and a system of national inspection and performance tables were all designed to raise standards of performance in schools and to provide better public information on the system. They implied a lack of confidence in teachers' and schools' ability to serve the needs of all pupils and raise standards of achievement overall.

The years since these legislative landmarks have added further accountability layers designed to raise standards further. Since 1997, there has been an emphasis on national targets and a requirement on LEAs to support school target setting and to intervene in schools where standards of performance remain a concern. The positive results have been a massive increase in public information about pupil achievement at all levels which has in itself acted as a lever in raising standards of attainment.

However, there have been negative effects for assessment. The growing emphasis on end-of-Key Stage tests as the basis of national targets and performance tables has resulted in teacher assessment taking second place. Teachers are often reported as regarding their role as teaching to the test with assessment driving the curriculum. Teachers say that what is assessed is that which is easy to measure but that these assessments do not always support learning. The Group's initial considerations of the 14-19 Reform Group had revealed that two terms of teaching and learning were lost in every GCSE course, for summative assessment preparation.

The development of value-added information on individual schools is a laudable aim but current methodologies are flawed and the tables have so far produced little usable information for parents and the wider community.

A recent report by the House of Commons Committee of Public Accounts said: '*Performance tables of academic achievement take no account of factors external to the school, some of which can have a significant influence on performance. The National Audit Office Report showed how academic achievement can be adjusted to take account of the influence of external factors to demonstrate the difference that schools make to the performance of their pupils. It also showed the effect of the adjusted data can have on how schools compare with each other*'.

The Council believes that performance tables, even with additional data, are a flawed source of public information and that a review is needed of the overall information available on school and pupil information for parents and the wider public.

The accountability context is beginning to shift, however. International comparative data resulting from the PISA research suggests strongly that our current assessment arrangements are not serving pupils' needs. Despite assessing children more regularly than any other country studied, England has a tail of underachievement at 16 which places it 15th out of 16 countries.

In its proposals for a New Relationship with Schools (NRwS), the government has made a public commitment to more 'Intelligent Accountability'. Schools are being given greater responsibility for their own improvement and development with more support being given to them in managing and interpreting their performance data and in the way that they provide continuing professional development A new model of 'shorter, sharper' inspections giving greater weight to school self-evaluation will be piloted as part of the NRwS during 2004-5.

Key Stage testing remains, but some elements of it are being overtaken by the new national strategies. A 2004 teacher assessment pilot in 36 LEAs has resulted in a Government announcement that summative assessment will be conducted by teachers at the end of Key Stage 1. Targets will originate at school level rather than being imposed by the LEA.

The Tomlinson Reform proposals for assessment envisage the development of a series of diplomas tailored to the individual needs of pupils. Accountability for standards of teaching and learning will reside with the school and the other institutions with which it collaborates. The e-assessment agenda being led by QCA with the future provision of on-line tests can support further personalisation of learning.

The Council supports the Reform Group's belief that different assessment and accountability models are needed to support the new 14-19 curriculum, and furthermore believes that the policy context in which we are moving requires a new framework for both throughout 5-19 education. The nature of such a framework is the theme of the remainder of this paper.

Baseline principles for the GTC assessment work.

The Council's starting point for its policy development work on assessment issues was a series of underpinning principles that it agreed in November 2002:

- The primary purpose of assessment is to provide feedback to shape and develop the teaching and learning activities in which both teachers and pupils are engaged;
- The national assessment regime should encourage a classroom culture of questioning, interaction and reflection;
- Assessment *for* learning rather than *of* learning occurs when evidence is used to adapt the teaching to meet the needs of the pupils, or by pupils themselves to change the way they work at their own learning;
- Strengthening the practice of formative assessment produces significant, and often substantial, learning gains. Many studies show that improved formative assessment helped low-attaining pupils and those with learning difficulties more than the rest;
- Evidence from many studies reviewed by Black and Wiliam shows that formative assessment does raise standards. The evidence also showed that there is room for improvement in the way that teachers use formative assessment;
- Test results are less effective at helping pupils to learn than the advice teachers can offer based on formative assessment;
- The contribution of teachers' assessments should be valued and developed rather than reduced in favour of external testing;
- The priority currently given to summative assessment must alter in favour of enhancing the development of formative assessment skills by teachers;
- The dominating influence of short, summative, external testing draws teachers away from formative assessment;
- External tests should be made more helpful by better understanding the interaction of external testing and formative assessment.

The current assessment regime means that the collecting of marks for record-keeping for external purposes is given greater priority than the analysis of pupils' work to discern learning needs. There is a need to reduce the emphasis on grades and marking.

Publishing raw percentage aggregate scores for each school is not an effective means of demonstrating public accountability and may not best serve parents' wishes to know about their children's learning and development.

Defining assessment terms

However, before exploring assessment issues in more detail, it is important to be clear about the terms that are currently used. The title of this paper makes a clear distinction between *internal* and *external* assessment. However, if external and summative assessments are not synonymous, internal assessment should not be equated with formative assessment as teacher assessment is often conducted for summative purposes.

The uses of summative assessment include diagnosis, tracking, grouping, reporting, target setting, certification and selection. (James, 2004, presentation at GTC Seminar, 18 May)

A review of the purposes of assessment is long overdue. The Council does not accept the original approach taken by the government's Task Group on Assessment and Testing (TGAT), that summative assessment in the form of an external test and a teacher assessment result could serve all the purposes of assessment as set out above. The Council therefore calls for a review of the purposes of assessment to secure a better alignment between the information sought and actually obtained.

[1] Inside the Black Box, Black and Wiliam. London: King's College, 1998

Teachers' views on the current assessment arrangements

The GTC has gathered evidence from teachers at a series of seminars on assessment during the academic year 2003-04. The following comments are typical:

'*Put the focus back on learning…Learning to learn should be a key skill for students*'

'*Where does formative assessment start and summative end?*'

'*Assessment and league table pressures affect teachers. They de-motivate us and those we teach*'.

'*Teaching to understand or teaching to the test?*'

'*What would parents prefer to hear? Their child's grades or a series of comments on their levels of improvement?*'

'*Overall, assessment and performance tables are creating tension with networking and collaboration*'.

'*Teachers need to be given back their professional judgement*'.

'*…workload issues need addressing. Workload and teacher confidence has been knocked*'.

'*How do teachers get pupils engaged in their own learning and become part of the assessment process?*'

Although many common themes emerged, there was genuine uncertainty about how to strike a better balance between measuring and supporting learning.

Uncertainties about the best way forward also appear in the GTC's Teacher Attitude Survey 2004, carried out on behalf of the Council by NFER. The survey was based on a stratified sample of 10,000 teachers drawn from the GTC database of 407,462 registered and practising teachers. The response rate was 44 per cent. It asked teachers to say whether, in their school, '*there is an appropriate balance between using assessment to support learning and using it to measure learning*'. The greatest proportion of respondents reported that they believed the balance to be right (40 per cent) while 39 per cent of respondents felt that there needed to be a greater emphasis on using assessment to support learning. However, there were others who felt that the measurement of learning needed more emphasis in their school (5 per cent) and those who were unsure (13 per cent). Additional analysis related to sector revealed that '*primary teachers are more likely than secondary teachers to report that the balance…is about right… A larger percentage of secondary teachers (47 per cent against 37 per cent of primary teachers) report that there needs to be a greater emphasis on supporting learning.*' (NFER, 2004)

Obviously a large-scale survey may contain ambiguities that have not been resolved by any subsequent interview, just as comments made in face-to-face contacts only represent a snapshot of stakeholder perspectives. However, the overall result is uncertainty about what assessment stakeholders, and teachers in particular, want in terms of an assessment model for the future. The GTC therefore believes that any review needs to examine the added value of a greater degree of teacher judgement has on learning progress and achievement. The main evidence source of such a focus must come from the body of assessment for learning research.

Assessment for learning

The evidence of the Assessment Research Group is that assessment for learning '*focuses on how students learn*', '*is central to classroom practice*', '*fosters motivation*' and '*develops the capacity of self-assessment*'. The Council supports these principles.

The AfL research formed compelling evidence at all three GTC seminars that formative assessment is effective. Representatives from Oxfordshire at one of the seminars were clear about the impact on progress. A teacher presenter reported that before she started, her pupils were not interested in her language teaching but only interested in the grades that they received. As a result of her AfL approach, pupils were more '*confident, motivated and engaged*' and were developing '*transferable skills and were not so content-driven*'. They had become aware of the assessment objectives, grown in confidence as they assessed each other's work and developed regular targets for improvement with her based on their own self-assessment.

Teachers involved in the Oxfordshire research also benefited professionally through partnerships, opportunities for school visits and peer observation, regular professional seminars and the support of assessment consultants.

The AfL research already reflects considerable evidence of the positive impact of AfL on summative results; a quarter to a half GCSE grade per student improvement is cited, proof that teachers '*do not have to choose between teaching well and getting good results*'. (Black et al, 2003)

There are still challenges to be resolved before Assessment for Learning can be used as the basis for a future teacher-led assessment model:

- there is a tension between the 'need to improve' messages of AfL and the 'successes and failure' messages of summative assessment for pupils (Black et al, 2003);
- the government's interpretation of AfL focuses on teacher use of performance data as the basis of dialogue and target-setting with pupils rather than using the individual pupil's learning needs as a starting point;
- the findings of the EPPI Review of the evidence of reliability and the validity of assessment by teachers used for summative purposes (Harlen, 2004) suggest a degree of unreliability of teacher assessment when used for summative purposes.

If AfL is to be the main foundation of the assessment framework for the future how is it to help deliver high-quality formative assessment, and contribute to the provisions for summative assessment information and the demands of wider public accountability?

Test issues

Public debate on the role of National Curriculum tests and external public examinations often portrays them as an objective assessment of pupil achievement that has greater currency than teachers' own assessments. The Council believes that the assessment review referred to earlier needs to include the current role of the statutory tests and tasks that we are using and to ask how effectively the claim to objectivity stands up.

Research concludes that no assessment method is neutral. While Harlen's EPPI research reveals that teacher expectation of pupils can affect assessment outcomes, standard assessment can be subject to bias in the setting of questions that make the process more favourable to some pupil groups than others. Gipps stresses that tests or tasks used for summative purposes and which aim to elicit the individual pupil's best performance need to be:

- concrete;
- presented clearly;
- relevant to the current concerns of the pupil;
- in conditions that are not threatening.

(Gipps, Presentation at GTC Seminar, February, 2004)

Gipps concluded that pupils needed a mixed economy of assessment approaches to give them the best chance of producing their best work.

Research has found that high stakes tests can have a negative impact on student motivation, for learning and for curriculum and pedagogy, (Assessment and Learning Research Synthesis Group, ALRSG, Harlen/Deakin Crick, Eppi, 2002), findings that echo some of the seminar perceptions earlier.

There are a number of technical issues about the tests. They include the remaining problems of scaling with a pupil assessed at the bottom of Level 4 being nearer in terms of marks to the top of Level 3 than the top of Level 4, and the weak criterion referencing involved in the system of testing. There is also the problem that public demand that tests maintain consistent standards over time would require everything related to the tests to remain the same. In fact the tests are curriculum-linked and the context on which they are based has been subject to constant change. Even if that had not been the case, students have become better at taking the tests themselves. (Oates, 2004, Presentation at GTC Seminar, 18 May, 2004).

The Council believes that a review of assessment needs to start from what is required of the assessment system both for the purposes of data, which can drive learning/achievement, and for the purposes of public information.

The Welsh perspective: Recommendations from the Daugherty Assessment Review Group

The remit of the Daugherty Assessment Group in Wales was to review the nature, suitability and timing of the current statutory assessment arrangements at Key Stages 2 and 3; the need for the Key Stage 2 assessment arrangements to better support the transition arrangements for pupils going from primary into secondary school; and the nature and the use of data for meeting each of the purposes of National Curriculum assessment.

At both Key Stages 2 and 3, the Group was '*persuaded by the evidence and argument that the current statutory arrangements are not as well-matched as they could be to the purposes they should be serving*'.

In particular, secondary schools do not make use of statutory assessment from Key Stage 2; as secondary teachers report that they do not get the information in time to use it. At Key Stage 3 in Wales, the statutory test outcomes are too late to inform subject choice for Key Stage 4. Another factor around the use of statutory tests at both Key Stages is the reported narrowing of the curriculum, an issue regularly raised by teachers at all Key Stages in England.

Recommendations from the Group for both Key Stages are for national tests to be phased out and replaced by teacher assessment. At Key Stage 2, a widening of information to parents is recommended based on skills testing and a profile of learning skills to be reported to parents at the end of year 5. At Key Stage 3, teacher assessment is to be carried out by the middle of term 2 of year 9 for it to play into subject choice. The proposals for enhanced statutory teacher assessment are supported by recommendations for a range of support materials, moderation based on clusters of schools and the use of Inset days for assessment training. At Key Stage 3 it is recommended that secondary schools are accredited *'as having in place procedures to maximise the consistency of statutory teacher assessment in each National Curriculum subject'*. Overall, the use of the Individual Pupil Learner Data Project is recommended as assisting both in benchmarking the performance of schools and local authorities and in the use of statutory measures of attainment to set targets for individual pupils.

GTC: An agenda for change

The GTC believes that any review of assessment needs to look at a series of fundamental questions: When do learners need summative assessment information to make subject choices? When do parents need assessment information? When is comparative assessment information useful to learn more about the system? The GTC believes that a similar assessment review of Key Stages 1-3 is needed in England to complement the review of assessment involved in the remit of the 14-19 Reform Group.

The Council supports a fundamental review that would include:

- the uses and purposes of assessment;
- the role of the teacher in assessment-building on the evaluation of the Key Stage 1 pilots as well as that of the Key Stage 3 English pilot - again developing the role of teacher assessment;
- the role of Assessment for Learning through the National Strategies and in LEAs working with the Assessment Reform Group action research;
- the role of national tests in the assessment system.

The GTC recommends that:

The government should conduct a fundamental review of statutory and non-statutory assessment at Key Stages 1-3 led by a reform group working similarly to the assessment part of the 14-19 Group;

The review of Key Stages 1-3 assessment should also consider the most effective ways of making formative and summative assessment information available for a range of audiences. This should include a review of the future role of published performance tables;

The GTC also supports the recommendation of the House of Commons Public Accounts Committee that the government needs to carry out a review of the best ways of making formative and summative assessment information available to parents, which would include the future role of performance tables.

(This discussion is referred to again in the later section on the New Accountability Framework.)

Recommendations could then feed into the 14-19 Reform Group recommendations for the future accountability framework.

The rationale and further recommendations that make up the Council's agenda for change in terms of pupil assessment are organised in the next four sections:

- A model for formative assessment;
- The model of summative assessment;
- The role of ICT;
- A new accountability framework.

A model for formative assessment

The Council believes that the Assessment for Learning (AfL) model being developed by the Assessment Reform Group (ARG) in conjunction with schools and LEAs should be the basis of teacher assessment. The GTC believes that the government needs to invest in AfL further through the National Strategies. The AfL model needs to develop the pupil as the learner who takes a role in his or her progress, developing the capacity for self-assessment and peer assessment. It needs to include more knowledge for the pupil and the teachers about '*learning to learn*' and the different kind of learning approaches that work effectively with individual pupils in a variety of contexts. This is particularly critical if AfL is to have any impact on the achievement with those pupils having the greatest learning difficulties.

However, the Council also believes that any developed model of AfL should not be used by teachers for summative purposes as this would distort its formative functioning. Instead it should be used as the basis for developing and training teachers in an enhanced model of formative assessment. In addition, by separating formative and summative assessment, the Task Group on Assessment and Testing (TGAT) illusion that one assessment can fit all purposes is avoided. The government's current interpretation of AfL suggests that summative performance data is equally helpful in determining fine-grained next steps in learning as data derived from formative interactions between teachers and pupils. This is regarded by the ARG as a misconception of AfL, which is not supported by any evidence. The government's interpretation of AfL with its top-down use of performance targets as a starting point could deter all pupils but, in particular, could deter those pupils with learning difficulties as well as not providing an appropriate learning environment.

The GTC is also convinced that a model of teacher assessment that fulfils both formative and summative functions constricts rather than enhances pupil learning and imposes unacceptable extra burdens on teachers. Even with the sources of professional support being proposed by the Daugherty recommendations in Wales (see Appendix), it could be an unmanageable model for teachers.

However, the GTC does support the development of assessment networks across LEAs with an AfL and formative assessment focus, similar to the arrangements in Oxfordshire as described in the AfL section of this paper.

The GTC recommends that:

- the Government should continue to invest in assessment for learning through the National Strategies working with the Assessment Reform Group. It must ensure that LEAs can support schools' and teachers' capacity to conduct assessment effectively, and should aim to revitalise learning in the process;
- that Government should develop a teacher assessment model that focuses solely on the formative assessment function to avoid distortion and prevent the imposition of unmanageable extra workload on teachers;
- investment in assessment for learning should include provision for LEAs to develop local assessment networks and to develop professional knowledge.

The Council believes that teachers' knowledge and experience of assessment processes needs to be a more prominent part of the professional standards framework at QTS, induction and continuing professional accomplishment. The GTC would be keen to work with relevant partners in this area of work.

The GTC also supports the direction of the development of work being carried further by the DfES and QCA, and supported by the 14-19 Reform Group on developing specialist roles for teachers in relation to chartered examiners and assessors. The Council is currently developing its Teacher Learning Academy (TLA), a framework for recognising and accrediting a range of teacher professional learning, and is keen to develop alignment of the chartered examiner/assessor initiative with the TLA.

The GTC supports the role of Chartered Assessors and/or examiners to lead the development of assessment communities at schools level. These teachers would have expertise in formative and summative processes and in secondary schools of involvement in public examinations. They would act as the development focus in the school on all assessment processes and might support interested but less experienced teachers. The GTC would again be keen to work with the DfES and QCA to integrate such specialist roles into teacher career paths and other parts of the professional standards framework.

The GTC recommends that the Council:

- works with the TTA and other partners to develop teachers' assessment skills as part of the professional standards underpinning QTS, induction and continued professional development;
- works with the DfES and the QCA to ensue that specialist expertise in assessment is reflected in teachers' career paths and with other parts of the professional standards framework.

A model of summative assessment

As the Council stressed earlier, it strongly urges the government to review all aspects of Key Stage 1-3 assessment including the current role and purpose of summative assessment and that of national tests and tasks. The GTC's recommendations in this section are made in this context and that of the need for a review of the summative assessment information available for a range of audiences and purposes, particularly the role of published performance tables.

The GTC proposes a three-strand summative model:

- a national bank of assessment activities/tasks to be administered with individual pupils/groups or classes at a time determined by the teachers;
- volunteer schools acting as nationally-accredited marking centres;
- national cohort sampling.

A national bank of assessment tasks to be administered with individual pupils/groups or classes at a time determined by the teachers:

The Council has been conscious that any proposals it makes on assessment should afford teachers the opportunity to hone their teaching and assessment expertise whilst not increasing workload, which is detrimental to teacher morale and well-being and thus, ultimately, to pupil learning.

Consequently, the GTC proposes that the QCA and awarding bodies develop banks of summative activities and tasks to be deployed by teachers as robust forms of summative in-time assessment for National Curriculum and public examination/certification purposes. Chartered Assessors/ Examiners should be involved in developing these banks. External marking of these tasks would be the norm with some schools choosing to function as accredited marking centres.

These activities/tasks would be a source of information on pupil achievement at a particular point, which could be summarised for the school, the LEA, parents and to pupils themselves during the Key Stage. Schools could use the information, for summative purposes such as pupil grouping and by schools, pupils and their families for subject choice as well as contributing to certification. The information from the activities/tasks could build towards an overall summative outcome at the end of the Key Stage, which would be crucial in contributing to learning decisions related to the next Key Stage. This would be particularly important at the end of Key Stage 3, when formative and summative information would contribute to the planning and reviewing processes leading into 14-19 course take-up.

End-of-Key-Stage summative information would also form the basis of the sampling process of pupil cohorts recommended to take place at the end of each Key Stage. However, for the reasons put forward in the accountability section, the GTC opposes the information being used for published performance tables.

In future, teachers and pupils nationally need a more effective means of providing more regular and personalised information on what has been learnt and achieved in the course of the key stage while losing the sense of being involved in a process of perpetually preparing for the test. The activities/tasks would place decisions about the timing and nature of summative assessment back within the professional judgement of the teacher without distorting or skewing the nature of AfL and teacher assessment. This would provide another complementary source of formative information for teachers and, as a summative source, be better integrated with teaching and learning.

Volunteer schools acting as nationally accredited marking centres:

The Council believes that schools should have the option to be accredited to assess summative activities/tasks for National Curriculum and 14-19 certification purposes. The role of Chartered Assessor/Examiner would be critical in the process of a school being accredited by the National Assessment Agency. Accreditation could also reflect a school's strengths in formative assessment and assessment for learning and it having developed a sense of creating an assessment community. Obviously such models could be very different in primary and secondary schools and in secondary schools, particularly large institutions and those with links to other schools, colleges and work places post 14,

there could be a role for subject assessment accreditation which could be wider than an individual institution.

The GTC recommends that:

- banks of summative activities and tasks should be developed externally, involving QCA, awarding bodies and chartered assessors/examiners and be administered by teachers.

The National Assessment Agency should ensure that the role of the Chartered Assessor/Examiner is pivotal to any consideration of the proposal for accrediting schools to assess summative tests and tasks.

National cohort sampling

The GTC believes that these testing activities should be de-coupled from the focus on the end of the Key Stage so that summative assessment data can better contribute to learning at the point it is needed.

End of Key Stage testing has undoubtedly provided valuable information for the system such as knowledge on the dip in achievement between Key Stage 2 and 3. However, this could be reached by means other than testing every child at the end of the key stage. Through representative cohort sampling, the system can continue to gather end of key stage assessment information for the purposes of year on year comparison and for analysis of trends and strengths and weakness.

Accordingly, the GTC proposes that schools are required to provide summative data for monitoring national standards involving two-year cohorts on a rolling programme.
The Council believes that the summative assessment of individual pupils be separated from the collection of summative data to be used for national monitoring.

The GTC recommends that:

- summative data is important for monitoring purposes at LEA and national level. Schools should be recruited to a rolling two-year programme that monitors a sample of pupils.

The role of ICT

The GTC is keen to work with the DfES, QCA and other partners to develop the e-assessment agenda. The GTC proposes that it organises a seminar on e-assessment by the end of 2004 involving teachers and other key partners in order to take perspectives on how best e-assessment can serve teaching and learning.

The Council believes that the 14-19 agenda is an appropriate place to begin the development of the provision of online tests linked to personalised learning and the longer-term agenda of building units of credit towards diploma certification. However, the GTC supports the development of online testing at all Key Stages and proposes that such an agenda should be linked to its recommendations for the provision of banks of summative assessment tests/activities and for accrediting schools to carry out summative tests.

The use of ICT should also support further improvements to the administration of summative assessments such as adaptive testing, electronic moderation of test data and new procedures for marking, such as individual markers being responsible for marking single questions relevant to their knowledge and expertise. This further underlines the need to support teachers in streamlining assessment processes.

The GTC proposes that:

- the Council should work with the DfES and QCA to develop the e-assessment agenda and should identify teacher perspectives on how this can best serve teaching and learning;

The QCA and awarding bodies should develop online tests as part of the creation of banks of summative assessment tests/activities, which could also be involved in the accreditation of schools to carry out summative tests;

the QCA and other partners should also support further improvements to the administration of summative assessments such as adaptive testing, electronic moderation of test data and new procedures for marking.

A new accountability framework

The GTC advocates a review of the information that should result from different kinds of assessment and how it should be appropriately articulated for a range of different audiences. Teachers need to be accountable to the pupils they teach but

one of the outcomes of AfL is the development of the pupil as a partner in teaching and learning process by involving the pupil in self-assessment.

Assessment for learning provides schools with a positive model for reporting to parents on a more ongoing basis. The GTC believes that changes to the assessment model would provide opportunities for schools to develop a new relationship with parents involving them in school and LEA-based assessment communities and in a dialogue with teachers that involves pupils and is based on narrative as well as on external test results

The Council also believes that the New Relationship with Schools, greater emphasis on school self-evaluation, better use of performance data by schools, and the development of the School Profile gives schools further opportunities to develop a new accountability relationship with pupils and parents. The Council believes that the review of assessment processes needs to include a review of the role of performance tables.

Performance tables can dominate school life, sometimes to the detriment of the learning opportunities of some pupils. Currently schools are being encouraged to take responsibility for their own development and to collaborate with other schools to improve learning opportunities for pupils. This is particularly key to the 14-19 agenda where collaboration is being promoted via the Pathfinder programme. The 14-19 Reform Group is due to report on their recommendations on the accountability framework to support their vision for future 14-19 curriculum, assessment and qualifications. These recommendations could reflect the need for accountability measures to better support local collaborative provision.

The GTC believes that performance information published in a more localised context would be more appropriate than national performance tables in order to reflect the personalised learning agenda as outlined in this paper. The Council supports groups of schools and colleges using data to improve learning across their boundaries and published data being available as a guide to provision available in the local area.

The GTC proposes that:

- changes to the assessment model should provide opportunities for schools to develop a new accountability relationship with parents which is based on a richer dialogue than external test results;
- the New Relationship with Schools, the greater emphasis on school self-evaluation, better use of performance data by schools, and the development of the School Profile should be promoted as a means for schools to develop a new accountability relationship with pupils and parents;
- the future of performance tables should be part of the government's review of the most effective ways of making formative and summative assessment information available for a range of uses and audiences.

The GTC has made a series of recommendations in this paper to work with other partners to develop this agenda for change. Its final recommendation is the role that the Council could take in brokering further development on assessment models. The Council has worked with teachers and LEAs to develop local models for continuing professional development and will be working in new groups of LEAs from autumn 2004 onwards.

The Council would be interested in working with teachers in a small number of these LEAs to develop the teacher assessment model it has outlined further, and to stimulate the creation of networks of expertise on assessment.

The GTC recommends:

- the Council will propose to the DfES that it works with a small number of schools and teachers on LEA CPD projects to develop the teacher assessment model further, and to encourage the creation of local assessment networks of expertise.

This paper is extracted from the GTC Council paper of June 2004 which was presented to the Secretary of State for Education and Skills as formal advice in October 2004.

Perspectives on Pupil Assessment

Publication date: November 2004

Publication code: P-NRCP-1104

Birmingham: Victoria Square House, Victoria Square, Birmingham B2 4AJ.
London: Whittington House, 19-30 Alfred Place, London WC1E 7EA.

General Teaching Council for England

0870 001 0308 info@gtce.org.uk www.gtce.org.uk