



ASF Working Paper 4

Case Studies

ASF Working Paper 4

Introduction

The Assessment Systems for the Future project aims to clarify thinking about the potential and the challenges of summative assessment by teachers. Working Paper 1 offers an analytical framework that can be applied to any context in which teachers assess, for summative purposes, their own students' work. Working Paper 2 summarises the evidence from two systematic reviews of research on aspects of using teachers' judgments for summative assessment. Working Paper 3 sets out the project's arguments for what is needed if summative assessment by teachers is to provide valid and reliable evidence of learners' achievements. This Working Paper seeks to illustrate how that analytical framework and those arguments can be applied to some examples of current policy and practice in a range of educational contexts within the UK. Applying a common framework to contexts that have previously been seen as distinct from each other shows there are generic issues that need to be taken into account if the full potential of summative assessment by teachers is to be realised.

Contrary to the impression sometimes given about the dominance of tests and examinations in the education systems of the UK, there is already a range of experience to draw upon of teachers having significant roles in summative assessment. There are therefore many examples that could have been included in a set of case studies such as this. These six case studies range widely across the education systems of the UK and across phases, from early years to vocational education. But there are many more contexts that could be analysed and discussed in similar terms, including the use of summative assessment within schools and colleges as distinct from teachers having a role within externally designed and managed systems.

The case studies have been written by members of the project's core group. In several cases the author's perspective comes from his or her close involvement in developing the assessment policies in question. In others, the author's perspective is that of an academic or professional with expertise and experience in the context referred to.

Each case study sets the scene with a brief overview of the context. That is followed by a summary of the main 'design features' of how teachers' assessments are used in that context. The contributors then analyse and evaluate that usage in relation to impact and practicability as well as validity and reliability. The final section of each case study highlights issues arising in implementing the system in question. In some cases this draws on the evidence that has accumulated over a considerable period; both the GCSE and NVQs were introduced in the 1980s. Most other cases – assessment in Scottish and Welsh schools, the Foundation Stage Profile in England – refer to systems that are only now being implemented. In one case – Key Stage 2 assessment in England – the final section considers an alternative system to that which is currently in place.

It is not the intention of case study authors to offer a full and authoritative review of the assessment policies and practices to which their commentary refers. Rather is it to

stimulate a wider debate about what might be involved if we are to develop to the full the contribution that summative assessment by teachers can make to assessment systems for the future.

Contents: Case studies

England

1. Foundation Stage Profile
David Bartlett
Co-ordinator for Assessment, Birmingham Children and Young People's Services
2. Key Stage 2: end-of-stage assessment
Paul Newton
Principal Assessment Researcher, Regulation and Standards Division,
Qualifications and Curriculum Authority
3. National Vocational Qualifications
Kathryn Ecclestone
University of Nottingham

Scotland

4. Assessment 3 to 14
Carolyn Hutchinson
Head of Assessment Branch, Scottish Executive Education Department

Wales

5. Key Stage 2: end-of-stage assessment
Richard Daugherty
University of Wales, Aberystwyth

England, Wales and Northern Ireland

6. General Certificate of Secondary Education
Gordon Stobart
University of London Institute of Education

1. England: Foundation Stage Profile

David Bartlett

Context

The foundation stage curriculum for England covers the nursery and reception years (four and five year olds) and consists of six areas of learning:

- Personal, Social and Emotional Development;
- Communication, Language and Literacy;
- Mathematical Development;
- Knowledge and Understanding of the World;
- Creative Development;
- Physical Development.

Progression within each of these areas of learning is defined by stepping stones that reflect development towards the early learning goals for the foundation stage. The early learning goals define the expected achievements for children in each of the areas of learning by the time that they reach the end of the foundation stage at the age of five.

The Foundation Stage Profile (introduced as a statutory requirement in 2003) presents the early learning goals in thirteen scales covering the six areas of learning. All areas of learning are assessed, reflecting the emphasis placed on the breadth of the curriculum in the foundation stage and the equal importance of each of these areas.

Design Features

Use

The main purpose of the Profile is to summarise children's achievements by the end of the foundation stage. However, it is also intended that it be built up gradually over time (predominantly during the reception year), with children's achievements recorded periodically, as and when they achieve particular points on the scales.

The Profile is designed to serve a number of purposes. Its day-to-day use is for monitoring children's progress, with practitioners using the information that gradually becomes available to adjust their planning and inform on-going learning and teaching. Second, as the Profile builds, it provides a framework for a dialogue with parents and for reporting to them. Third, by the end of the reception year (i.e. the end of the foundation stage) the Profile also provides a summary of each child's progress in relation to the early learning goals. These summaries can be used for reporting to children's next teachers as they move into key stage 1. In addition, these assessments are submitted to the local authority and in turn the assessments must be submitted to the DfES. Individual assessments are only submitted to the DfES for children in Sure Start areas. Beyond this, only summary assessments are submitted, so that the DfES or other agencies with access to national data cannot identify the assessments for individual pupils and schools. While the information from the assessments can therefore be used by schools and local authorities for monitoring school performance, the information can only be used by the DfES for monitoring overall national standards and standards in Sure Start areas for a 10% national sample.

Criteria used and judgements made

There are thirteen scales that make up the Profile (three for Personal, Social and Emotional Development; four for Communication, Language and Literacy; three for Mathematical Development; and one each for Knowledge and Understanding of the World, Creative Development and Physical Development). There are nine points in each scale, the first three points drawing on the stepping stones, points four to eight being the early learning goals themselves, with the last point in each scale reflecting progression beyond the early learning goals. Assessments are made with reference to these points and accompanying exemplification within the Profile Handbook.

The points within each scale are descriptions of achievement, with assessments being made from evidence arising from children's on-going learning in relation to these, rather than in relation to task specific criteria. Judgements reflect children's progression towards and achievement of these criteria. They are made on the basis of evidence from children's day-to-day learning and not from specified tasks. However, the Profile Handbook does provide outlines of activities that practitioners can use on an optional basis to provide additional evidence of children's progress.

Guidelines/training

The FSP Handbook and accompanying CD ROM provide the core guidance for understanding and using the Profile, although additional materials have also been published by the DfES and other materials have been produced by individual local authorities and regionally by clusters of local authorities. Training for the Profile is developed and delivered locally although this is done with reference to both nationally and locally produced materials.

Moderation of teacher judgements

The Profile Handbook contains a model for local authority moderation, involving both visits to schools, and other settings where the Profile is in use, and cluster meetings of schools. The model provides guidance about the form that visits and cluster meetings could take, although it does not specify these in detail or provide materials to be used. The model consists of a three year cycle covering the six areas of learning for the foundation stage, with a focus on two of the areas within each school year. It was intended that all settings should receive moderation visits within the first two years of the Profile being introduced, with visits to 25% of settings being made in each subsequent year.

Properties

Validity

The construct validity of the Profile is potentially high, given that it covers all of the areas of learning for the foundation stage and includes all of the early learning goals. Judgements are made as to whether children have achieved particular early learning goals using a wide evidence base from on-going learning and teaching, over time and over a range of contexts and learning experiences. However, the curriculum guidance for the foundation stage emphasises that learning should be predominantly child initiated and play based. The degree to which the assessments made by practitioners (both teachers and teaching assistants) properly reflect children's achievements of the early learning goals will depend not only on the coverage of curriculum content, but also on the effective implementation of the curriculum processes, through child

initiated activities. To the extent that some of the curriculum content, as well as the curriculum processes, may be unfamiliar to practitioners, it will take time for the foundation stage curriculum to become embedded and consequently it will also take time for the potentially high construct validity of the Profile to be realised.

Impact

The Profile is intended as a summary of children's achievements by the time that they reach the end of the foundation stage, serving the same purposes as end of key stage assessments, i.e. reporting to parents and children's next teachers and providing summary information for school evaluation purposes. However, it also serves other purposes. If judgements for the Profile are made gradually in the way intended, it provides information about children's progress that can be used to inform planning and adjust on-going teaching, prompting practitioners to develop the ways in which they are delivering the curriculum. It also promotes a summative assessment model where judgements draw upon the rich assessment evidence arising from on-going learning and teaching. Because this evidence arises from assessments that have a short-term formative purpose, summative and formative assessment processes are brought together as the Profile is built over time. It is the intention that children should not be aware of the summative assessments being made. If the Profile is not built in this way but completed at the end of the year, these positive consequences are lost. If the Profile is used in the way intended, consequential validity for reception practitioners is potentially high.

Profile assessments cannot be used to make comparisons between schools in the same way as for National Curriculum assessments and examinations, as aggregated results are submitted to the DfES by local authorities, within which the assessments for specific schools cannot be identified. Nevertheless, local authorities are still able to produce comparative information for schools and schools' assessments can be compared with national data at the time of inspections. Accountability judgements are therefore possible and these could have negative consequences for the Profile assessment processes. However, this will depend on how the information is used and the extent to which schools can use it for self-evaluation, rather than seeing themselves as responding to external judgements.

Reliability

A further characteristic of the Profile is that all practitioners in a setting (both teachers and support staff) should be involved in the assessment of children and in making shared judgements. This constitutes a process of moderation between practitioners through which internal reliability can be developed. External reliability is addressed through moderation meetings that bring together practitioners from different settings to share their assessments and develop consistency in their interpretation of the early learning goals. Since these assessments can only be understood in relation to the contexts within which they are made, moderation meetings also provide a vehicle for practitioners to share the ways in which they are implementing the curriculum. Moderation processes are therefore not only a way of developing reliability, but also a powerful vehicle for professional development, through the sharing of practice.

Practicability

If the Profile is not built gradually in the way intended, practitioners are faced with the unmanageable and time-consuming task of summarising all of their judgements towards the end of the reception year for all of the children in a class. While this may

serve the purposes of reporting to parents and next teacher and provide the required information for external data collection, it would clearly be an exercise of limited usefulness. If the Profile is built gradually over time, manageability is in the hands of practitioners. It does not require any evidence to support judgements over and above what would normally become available from on-going learning and teaching, although practitioners need to know the curriculum and be able to articulate the judgements they are making.

The scope of any system utilising teacher assessment is potentially problematic because over time large amounts of information can be generated that can be difficult to summarise and interpret if records are paper based. The electronic record for the Profile (the eProfile) is an effective solution to this in that it summarises assessments in a way that makes them manageable and accessible so that they can be used to inform planning for classes, groups and individual children. While its use is not a requirement, without it practitioners can experience manageability problems and much of the power of the Profile for monitoring and planning is undermined.

Issues arising in implementation

As a statutory assessment, the Profile is an external requirement for assessment that is unique in the English primary system, not only because it is based entirely on teachers' assessments but also because it requires assessments across all areas of the curriculum. The experiences gained from its development and implementation provide a number of valuable lessons for any further introduction of summative assessment systems utilising teachers' own assessments. Some of these lessons are considered below.

Implementation and support

The Profile was introduced hurriedly part way through the 2002-2003 school year with teachers expected to summarise their first assessments by the end of the school year. This did not allow teachers time to familiarise themselves with the Profile, made it difficult for local authorities to organise the first round of training and in any case was inappropriate, given that the Profile is designed to be built gradually over the reception year. Experience with the Profile indicates that any change to a system based on teachers' own assessments requires time for the system to become established. It is only now that the Profile is becoming properly embedded and beginning to impact on learning and teaching in the foundation stage. Clearly, any substantial change in statutory assessment processes requires an understanding on the part of policy makers of the changes themselves and properly planned implementation.

Moderation

In the three years since the introduction of the Profile, there has been considerable variation across the country in the extent to which moderation processes have been developed. In some local authorities moderation is well developed and in addition, in some parts of the country, there are annual regional conferences for moderators to develop consistent approaches and interpretation of the Profile scales. In other areas moderation processes have not been developed.

Three years after the introduction of the Profile support is now being provided by the National Assessment Agency (NAA) in order that moderation processes are implemented across all local authorities. Despite the moderation model contained within the Profile Handbook, the slowness of the DfES and the NAA to encourage moderation processes nationally suggests that policy makers and government agencies do not appreciate the importance of moderation in any summative system utilising teacher assessment.

Impact on other stages of education

Year 1 teachers have experienced difficulty in interpreting and using Profile assessments, as the Profile reflects achievement of the early learning goals rather than being expressed in terms of the National Curriculum attainment targets that year 1 teachers are familiar with. However, this is a consequence of the foundation stage curriculum being expressed in terms of early learning goals, rather than a consequence of the Profile itself (which was written to reflect the curriculum). In addition, the child initiated learning of the foundation stage contrasts with the relatively more formal approach that can characterise teaching and learning in key stage 1. The combination of these two factors may have resulted in year 1 teachers making little use of the Profile information and attributing little validity to the Profile assessments that they receive. However, both local and national initiatives are being implemented to address these transition issues, with teachers developing their understanding and use of Profile assessments and the foundation stage approach to learning, as children move into key stage 1. A general conclusion from this is that where changes are made to the summative assessment processes for any part of an education system, it is essential that the potential impact on other parts of a system be given proper consideration.

2. England: end-of-stage assessment at Key Stage 2

Paul Newton

Context

At the end of year 6, almost all 11-year-old pupils in England are assessed in English, mathematics and science to provide summative statements of attainment. Two parallel assessments occur, one based on tests and one on teacher assessment.

Test results are used for a variety of purposes, with potentially high stakes for: policy makers (national monitoring); schools (local competition, organisational intervention, resource allocation); teachers (performance evaluation); and pupils (secondary school placement). Both test and teacher assessment results are used for a variety of other purposes, with lower stakes.

Design features

Both test and teacher assessment results report primarily in terms of national curriculum subject levels, which characterise attainment from year 1 to year 9 using an eight level scale.

Teacher assessment involves judgements of ‘best fit’ between evidence of pupil attainment and specified level descriptions. The nature of the evidence base, and exactly how judgements ought to relate to it, is not specified. Judgements are made for a small number of attainment targets per subject, from which overall subject levels are derived. Moderation is officially encouraged, but not required.

Each test comprises at least two papers, lasting around 2 hours in total. They are developed by test development agencies, using question formats which range from selected response, to both short and extended constructed response. Procedures for linking test standards between years enable thresholds to be established which divide the mark scale into national curriculum level bands. Tests are marked externally, by teachers who are trained to apply the mark scheme, and whose performance is standardised, monitored and graded.

Properties

The defensibility of any set of assessment results can only be judged in relation to the purposes for which those results will be used. Given the multiplicity of high and low stakes uses associated with national curriculum assessment results, and the fact that there is no clear specification of purposes, this is problematic.

Validity

The tests are designed to represent national curriculum programmes of study, as broadly and fairly as possible. However, the relatively small number of questions limits coverage, and there are certain aspects of each programme which are very hard to assess through the available question formats which do not include extended performance assessments.

The fact that broadly similar kinds of test question appear each year, on broadly similar content areas, raises questions concerning the extent to which improvement in the national profile of attainment represents: either an increase in the general quality of teaching, *per se*; or an increase in the ability of teachers to prepare pupils for the specific demands of the national curriculum and its tests. This is a contentious debate and there is not a great deal of evidence on the matter.

Teachers have much more, and a much broader range of, assessment evidence available to them. However, without a systematic approach to evidence generation, evaluation, recording and aggregation, we should not assume that teacher assessment results are necessarily more valid than test results. Both bias, whether intentional or not, and ineffective aggregation of evidence are significant threats to the validity of teacher assessment judgements, given present practices.

Reliability

Test papers tend to fare reasonably well when judged according to traditional indices of reliability, such as the ‘Cronbach alpha’ statistic. But a more important question is whether a pupil would have been awarded the same level if s/he happened to have taken a different version of the same test. We have very little evidence of this kind of reliability, although some have suggested that as many as 30% of pupils might receive a different level.

We have virtually no evidence of the reliability of teacher assessment at key stage 2. It could be better than for tests, since teachers notionally have a lot more assessment evidence on which to base their judgements. But we have no evidence of this. It seems likely that some teachers’ judgements will be considerably more reliable than can be expected from test results, while some teachers’ judgements will be considerably less reliable. This raises important issues of fairness in the use of results. It also seems likely that teachers, on the whole, will be less likely to make serious misappraisals, as can occur with tests when (for example) pupils suffer from severe test anxiety or when markers make major addition errors.

One aspect of reliability, known as comparability, concerns whether the standard applied for one group of pupils is the same as that applied for another. Here again we have little firm evidence in relation to key stage 2. However, based on common sense and evidence from other contexts, we might assume that comparability of standards between schools – crucial for school comparison – is likely to be higher in relation to test results than teacher assessment results (especially when, as is usually the case, no cross-school moderation is undertaken).

Practicability

National curriculum tests are relatively easily implemented by teachers, since they are both designed and marked externally. However, this comes at a large financial cost to central government. Problems also exist in recruiting sufficient numbers of markers to undertake the exercise; and the extremely rapid turn-around time can cause problems for the External Marking Agency.

Particularly following the 1993 Dearing review, the process of reaching teacher assessment judgements has become more practicable for teachers since they are only

required to make a small number of holistic judgements for each pupil at the end of each year. However, this practicability may come at a cost to the quality of the teacher assessment judgement. The more an individual teacher engages in a systematic process of generation, evaluation, recording, aggregation and moderation of assessment evidence, the higher the technical quality of her/his assessment judgements are likely to be. Yet, unless there are compensating benefits such as the use of such assessments for formative purposes, they will also be correspondingly less practicable.

Impact

The high stakes which presently attach to test results carry with them two main risks. First, too much attention may be given to questionable pedagogical strategies which impact directly on test results but also threaten broader aspects of educational attainment. Second, too little attention may be given to positive pedagogical strategies which can impact both on broader aspects of educational attainment and also on test results although not necessarily directly or immediately.

The first includes the risk that ‘drilling’ pupils in strategies which enable them to perform well in national curriculum tests may not effectively inculcate the robust, generalisable and useful kind of understanding that the tests are meant to indicate. It also includes the risk that undue attention will be given to areas of the curriculum which are tested (English, mathematics and science) at the expense of other areas or that undue attention will be given to areas of the tested subjects that are well represented in the tests at the expense of less easily assessable aspects.

The second includes the risk that the dominant discussion of assessment in terms of summative tests will orient teachers away from the importance of developing skills of formative teacher assessment in the classroom. It also includes the risk that teachers, managers nor policy makers will recognise the development of summative assessment skills by teachers as a fundamental aspect of teacher competency.

An example of an alternative system

One way to improve the quality and impact of national curriculum assessment at key stage 2 might be to rationalise the relationship between the uses to which results are put and the components of the system that generate those results: either by eliminating certain purposes entirely (e.g., national monitoring, school competition, or teacher evaluation); or through re-designing the system such that specific components are tailored to satisfy specific purposes. What follows here is a three-component example of the latter approach.

National assessment

To monitor trends in the national attainment profile of successive cohorts of 11-year-olds, the preference should be for an assessment component which:

- tests under low stakes conditions (so that teachers have no incentive to ‘teach-the-test’, and so as to ensure the security of test questions)
- tests different pupils on different blocks of questions (so as to sample the assessed material thoroughly using a full range of question types)

- tests a representative sample of the cohort rather than the entire cohort (so as to reduce costs and to ensure manageability)

In monitoring even short- to medium-term trends it is wise only to focus upon subject areas (and aspects of those areas) which are least likely to change significantly in relevance over time. Change in the educational significance of assessed material introduces unavoidable ambiguity into the interpretation of trend lines. Note that the monitoring tests would not aim to represent national curriculum programmes of study as a whole but only those aspects that were deemed central and least likely to change significantly in relevance over time.

The need to run complex statistical models might limit the range of question topics and formats which could be addressed (to some extent) but this is a compromise which might need to be made to enable aggregate results which could form the basis of reliable trend lines.

An example of such a system is the long-term trend National Assessment of Educational Progress, which periodically assesses attainment in the core aspects of reading and mathematics, in the USA.

School assessment

National curriculum test results are presently used for a variety of purposes based upon the principle of school comparison, including local competition (between schools), organisational intervention (for failing schools), resource allocation (to support areas of strength and/or weakness). School comparison requires that all schools be subject to the assessment and that the assessment process ensures a high degree of comparability. It seems likely that measures of pupil attainment for school comparison purposes are best supported by tests similar to those presently in use. Sole reliance on teacher assessment would present a threat to the comparability of results between schools, even if only a perceived threat (although probably an actual one as well).

However, even retaining the present suite of tests in English, mathematics and science, there are good reasons to make some significant changes. A first suggestion would be to run the tests biennially rather than annually. The national monitoring assessment could be run in the intervening years. On balance, this would reduce the burden of testing on primary schools.

A second suggestion would be to convert results to scale scores rather than to national curriculum levels (standardising the national distribution of marks in each subject to the same mean and standard deviation each year). This would elevate the status of teacher assessment results, which would become the only mechanism for reporting national curriculum levels. It would also entirely eliminate the possibility of level setting error, which is a risk for each test each year at present. Trends over time would now be monitored through a separate instrument. Comparisons between schools, within years, do not require strict comparability of standards over time. Not setting levels on national curriculum tests would also mean that the test development agencies would be far less constrained and would be free to evolve the tests over time to keep pace with advances in curriculum, pedagogy and assessment in ways that are not possible at present.

A final suggestion would be not to report individual pupil scale scores back to pupils. Individual scores are not reported to pupils who participate in national monitoring exercises since they are insufficiently reliable at this level. This would further elevate the status of teacher assessment as not simply the only mechanism for awarding levels but also the only mechanism for producing pupils' results at the end of key stage 2. It would also be sensitive to critics of tests who note that providing norm-referenced results to pupils can be harmful for those who repeatedly reside at the bottom of the distribution. Equally importantly perhaps, if there were no requirement to report pupil results to schools by the end of the summer term a greater investment in marking would be possible, resulting in more accurate results.

Of course, since the tests would still be high stakes for schools, this approach would not necessarily dispel the threat of 'teaching-the-test'. However, the more professional approach to teacher assessment discussed below might help to reduce this threat.

Pupil assessment

For many of the purposes we have in mind when assessing individual 11-year-old pupils subject level test results are far from being ideal. For example, for the purposes of progress tracking, diagnosis of educational needs, transfer and placement decisions (etc.) finer grained analyses are more useful. These are made routinely for teacher assessment judgements at the attainment target level.

However, the main problem with teacher assessment at key stage 2 is that it is not based on a coherent model of practice and, as such, cannot guarantee rigour. Also problematic at the end of key stage 2 can be the process of pupil transfer from primary to secondary school. This is exacerbated when, in particular, secondary teachers do not trust pupils' end of key stage 2 results (test or teacher assessment) and, effectively, start the assessment process from scratch. The transfer process has much potential to be improved and the development of an effective model of teacher assessment practice could be an excellent mechanism for achieving this.

One way of improving the transfer process might be to ensure: first, that both primary and secondary teachers were part of the same 'community of practice', sharing the same assessment standard; second, that not only the pupil and her result transferred, but also the evidence upon which that result was based; third, that the assessment process in year 7 takes off from where it ended in year 6, even to the extent of updating the same evidence base.

To achieve this assessment goal would require a number of key process features. Teachers, in collaboration with pupils, would be required to prepare – and to keep updated – folders of work in each subject which represented the 'latest and best' evidence of performance in relation to the national curriculum Attainment Targets. This would be akin to the approach presently adopted for school-based assessment in Queensland. Another core requirement, also characteristic of the Queensland approach, would be a commitment to local moderation. This would involve both year 6 and year 7 teachers from the area and might involve work from both years 6 and 7. The moderation process would emphasise quality assurance as much as quality control and would be undertaken well before final results were produced. It would

also emphasise a negotiation of the standard amongst peers, although the final decision for each pupil would always rest with her/his teacher.

Although the results would not have very high stakes, there would be a heightened sense of professionalism for all involved, since there would be a literal hand-over of folders (and professional responsibility) at transfer.

There would be no need to report overall subject results, and results might be presented at the attainment target level only. This would help preserve the integrity of the assessment process and it is not obvious what would be lost by not reporting overall subject levels.

Finally, this approach – which makes the assessment process central to the teaching and learning experience – is likely to be more supportive of the implementation of formative assessment although it would not in itself guarantee good formative assessment practice.

3. England: National Vocational Qualifications

Kathryn Ecclestone

Context

Introduced in 1989, National Vocational Qualifications (NVQs) were a radical and controversial overhaul of work-based qualifications. They were supposed to update and improve the quality of work-place training by reflecting the demands of employers for key competences in specific occupational roles. Designers of the new qualifications wanted assessment to be carried out by internal work-place assessors rather than ‘teachers’. The term ‘teacher assessment’ does not therefore resonate with the NVQ system, unless college teachers are also assessing candidates as part of a work-based programme.

Specifications of competence and criteria for assessing and accrediting them were created initially by industry lead bodies representing employers’ interests in different occupational sectors; these were reorganised as sector skills councils in the 2001 Learning and Skills Act. NVQs can be taken whilst working in a job or apprenticeship scheme, on a specific course in a college, a qualification and training programme offered by a training provider or a mixture of these formats.

NVQs were designed to be available from entry level to employment to degree and professional equivalent. In reality, the biggest growth areas for NVQs have been at level 2 in the qualifications framework for public sector workers in health and social care, retail and leisure industries and in the use of NVQs to accredit modern apprenticeships taken by school leavers.

The extent of employer demand for NVQs has been undermined both by the continued availability of non-NVQ qualifications and by the way that the vast majority of funding for NVQs has come from taxpayers. The goal of employers paying for NVQs after their initial introduction has not been realised. Without compulsion for public sector organisations such as Royal Mail and the NHS to offer NVQs, and significant amounts of public subsidy, NVQs would not have got off the ground .

Over the past 15 years, far from achieving the original goal of a radical overhaul through system-wide implementation of NVQs, England has developed an extraordinarily complicated and confusing system of work-based qualifications. There are thousands of vocational qualifications, hundreds of awarding bodies and thousands of providers including FE colleges, employers and private training organisations. In addition to being accredited by long-established, well-known and respected vocational bodies such as City & Guilds, NVQs can be accredited by unions, professional organisations, sector-specific bodies and small awarding bodies.

Design features

Specifications and their design

The *competence-based assessment* regime of NVQs is a particularly strong form of criterion-referencing which emphasises the authenticity and validity of work-place competences. Detailed specifications of competence in different roles, the range of contexts in which it must be demonstrated and the indicators of performance that show it has been achieved are used by assessors and candidates to assess and record achievement, or to set further targets. The goal of its designers was that accurate specification would enable assessors to assess validly and thereby reliably.

Sector skills councils are responsible for creating and updating the specifications to reflect employers' needs in different organisations within a particular sector, such as retail. An NVQ comprises units of competence that can be taken and accredited separately and as part of a whole qualification.

Summative assessment

Summative assessment is based on the internal assessment of practical competence by workplace supervisors in assessment roles, college teachers and other assessors, supplemented by externally designed unit tests of underpinning knowledge; some of these are accessed on-line as and when candidates are ready to take them. There are requirements for training in competence-based assessment for anyone assessing NVQs.

Assessment demands centre on what the learner (often also the employee) can do, and can be seen to do, in relation to the tasks required of them for competent practice. Detailed specifications of outcomes and assessment criteria promote and demand 'mastery' (i.e. coverage of all demands) as opposed to compensation and grading (i.e. where assessors can off-set poorer performance in some areas by better performance in others).

Candidates are required to show evidence of workplace competence in diverse forms, relevant to demonstrating mastery. These include: observation by supervisors and/or external assessors; written testimony by colleagues or managers; written assignments; practical tasks; oral feedback and testimony. There is a strong emphasis on assessment tasks being 'fit for purpose' and the validity of assessment as opposed to reliability. Candidates can repeat assessment tasks until they are deemed to be competent, producing assessment decisions of 'not yet competent' (working towards...) or 'competent'.

NVQs require learners to demonstrate achievement when they are ready ('readiness') Detailed help can be given in the form of formative guidance and feedback and, in some cases, repeated assessments until the candidate achieves the outcomes. NVQs are also supposed to be rooted in authentic, work-place contexts and assessed by people inside those contexts. Thus 'achievement' is defined in terms of demonstrated competence, while 'fairness' involves transparency of criteria and procedure, comparability/similarity of assessment tasks and contexts, and multiple opportunities to demonstrate the required competence(s). In NVQs, candidates can repeat tasks until they demonstrate competence, with as much guidance as necessary.

Formative assessment

Formative and summative assessment are synonymous in NVQs, where diagnosis, target-setting, review, recording competence becomes the learning (assessment) programme. Assessors gear their coaching and teaching to meet the next competence or target, according to how well trainees performed summatively in the previous one. A number of assessors might be involved in one trainee's programme: a supervisor at work, a colleague, a teacher/trainer in the college or training provider, the employer.

Some tasks cover a number of competences. Evidence gathered during learning and work activities is interpreted in terms of progress towards units of competence and the smaller tasks that comprise those units. The involvement of trainees in this process is extremely variable from one setting to another. In some it seems that assessors compile the portfolio for the trainees so that they are almost unaware of how their tasks at work lead to the NVQ! In others, the self-assessment, review and close engagement by trainees with designing their programmes of work and assessment, envisaged by the architects of NVQs, are more visible.

Assessment is also used to accredit prior learning (APL), where review of past performance and activities, and evidence from these, can be used to award units of NVQs without new learning being needed.

Properties

Validity

This is central to the design and specification of the units, competences and performance criteria. The NVQ was designed with this property as its central goal in order to ensure that assessment was fit for purpose, did not require unnecessary theory, was 'employer-led' and based on authentic, real-life tasks for specific occupations. However, validity rooted in the authenticity of work-based tasks is compromised by simulation of activities, for example in colleges, and by lack of access to the full range of occupational tasks needed to make up a unit of competence.

There are questions in NVQs about how far the diverse range of methods used as 'evidence of competence' reflects the learning being assessed. Specifically-designed tasks and tests are designed for learning and assessment as synonymous activities defined in the criteria and specifications. These form the 'syllabus' and set the goals of learning. How far these are wide-ranging, covering skills, attitudes and creative and critical thinking skills is questionable and varies among NVQs.

Impact

NVQ outcomes are used to judge the institution's or organisation's overall achievement rates, and where NVQs are funded by the LSC, these outcomes are monitored as part of national targets for the achievement of NVQs at different levels. Teaching and training focus on the summative outcomes but this as much a product of the strong criterion-referenced format and the emphasis on teacher (assessor) assessment as the way that outcomes are used for accountability.

A tendency to focus on a narrow interpretation of criteria and on performance rather than learning is not solely because of excessive accountability for summative results. It is also because of resource pressures to get trainees through the process as cost effectively as possible and the closely specified and prescriptive competences and criteria. It is not clear in NVQs how some of the positive consequences of greater freedom compared to external tests might apply; the pressure to pass the competences comes from other sources, not least the fusion of formative and summative processes and activities.

Reliability

It is important to note that 'standards' in NVQs mean 'occupational standards of competence' rather than its more familiar meaning of 'comparable standards' of achievement within qualifications or between centres and cohorts. NVQs trade national reliability and standardisation of assessors' judgements between centres and providers in order to privilege validity and authenticity. Awarding bodies therefore place more emphasis on verifying that procedures and guidance have been adhered to because of the sheer variety of practices, providers and programmes. However, awarding bodies do moderate and sample assessment decisions and so, although reliability is a feature of NVQs, it perhaps takes a different form than in other qualifications.

The chief source of low reliability in NVQs is the local, individual nature of interpretation of evidence by assessors. Institutions are required to carry out internal verification and moderation of procedures and outcomes, with some sampling of portfolios. Awarding bodies follow this up with annual visits to providers by a subject specialist and issue annual reports about the national standards of work in each area. QCA does not collect national data to compare centres and providers in terms of NVQ outcomes. Comparability of outcome and provider is therefore arrived at through tight specifications of tasks and procedures rather than by moderating assessment decisions.

Practicability

The running costs of NVQs are high because of the intensity of the assessment process and its individual focus. There is wide variation in quality of training underpinning the assessment process, in terms of time spent on training, individual reviews and portfolio building. Assessors also report spending a great deal of time translating the criteria, tracking the evidence in the portfolios and generally auditing and managing assessment and quality assurance processes for the awarding bodies in a complex and prescriptive assessment regime.

In addition, where providers offer different NVQs from different awarding bodies, the administrative and quality assurance costs can be very high: a large college, for example, might pay fees to 20 different awarding bodies in order to offer NVQs and other vocational qualifications. This also requires creating links between different quality assurance procedures for verification and moderation.

Issues arising in implementation

The concerns discussed here arise from assessors, candidates and awarding body officials interviewed during a recent, in-depth qualitative study for the Learning and Skills Development Agency [ref. here] Some of these concerns were also raised by Alison Wolf from her work with the design of NVQs in the late 1980s and her analysis of competence-based assessment (Wolf, 1995).

Understanding the assessment demands

One element in characteristics of validity and reliability is the enduring opaqueness of the outcomes-based, competence-based language of NVQs. While improvements have been made to many of the original specification documents, successive generations of learners and assessors have to learn the language anew, despite the efforts of awarding bodies to simplify the language and format of the standards. Assessors in the LSDA project reported the need to repeat information, to translate the ‘gibberish’ and ‘wordiness’ of the specifications and to wait for understanding to emerge after candidates have done one or two units.

Getting candidates through the requirements

The LSDA project showed very high levels of support and guidance in the process of identifying evidence. There are many examples of tutors and assessors interpreting awarding body specifications and criteria and providing simple translations of what they “really mean”. Sometimes these examples tread a very fine line between eliciting what the candidate ‘knows’ and leading them, even word for word, to articulate the desired answer. For example, there were numerous examples of workplace settings where such support was observed through ‘leading questions’ by assessors of a ‘good lad’ to help him through observations of his workshop practice and compile his portfolio evidence. Similarly in Social Care, assessors were observed asking leading questions to help candidates articulate what they (supposedly) already know and can do.

The practices of translation, support and direction revealed by the LSDA project have implications for the desirable interpretation and mediation of assessment procedures and practices at local level. There are also questions about the nature of the knowledge candidates have and how they acquired it. The project raises questions about the extent to which range statements and performance indicators can be said to ‘represent’ the reality of workplace competences if they are not recognized as such in the workplace, have to be ‘translated’ for workplace use, and if observed competences have to be translated back again into acceptable evidence statements.

None of this is necessarily inappropriate or unfair in itself since it might be argued that such practices are at the heart of professional judgements about the interaction between performance and competence that all assessors must make in different regimes. Nevertheless, there are problems about equity if they are not pursued uniformly: questionnaire data in the LSDA project suggest that there can be wide variations in the frequency and length of assessor visits. For example, while the most frequently reported timing of assessor visits amongst the NVQ-takers was 1-2 hours every 4-6 weeks, one reported that they saw their assessor once a week for 2-3 hours, while four reported that they saw their assessors only every 3 months or less *and* for

one hour or less. Some regulation of, and minimum recommendations for, such visits and levels of support seem to be needed.

‘Fair’ assessment

It is important to note that the emphasis on validity and authenticity, mastery and competence in NVQs led assessors in the LSDA project to have very different notions of what counts as ‘fair’ assessment in NVQs compared to other post-16 assessment regimes. This affects tutors’ and learners’ attitudes to what counts as ‘achievement’ and to the respective roles of formative and summative assessment. In contrast to NVQs, for example, formative assessment in Advanced Vocational Certificates of Education (AVCE) helps students improve their grades and this is integral to the educational ethos of the qualification. This takes the form of guidance on draft assignments and close attention to the criteria which students are encouraged to use in detail.

Opportunities for valid assessment

The LSDA study showed that what one might call ‘opportunities to verify’ vary greatly across work-based and college-based settings. In turn, this has implications for reliability of the standards of competence achieved. For example, small garages may not provide NVQ level 3 opportunities to conduct diagnostic work with the latest computer technology. Equally however, and somewhat ironically, well-resourced main dealers for leading carmakers do not always provide NVQ level 2 opportunities for basic repair – “clutches don’t go wrong on Volvos”. In Sport & Recreation small hotel leisure facilities can be very limited in the equipment available, and indeed in client activity, so simulation is often called for with another member of hotel staff acting as a client. Often therefore, ‘ways and means’ are found to observe and verify competences but this can impede progress and have negative effects on learners and assessors’ motivation.

A related issue with respect to ‘opportunities to verify’ is that candidates may not necessarily be in a position to gather relevant evidence. For example candidates already have to be in a supervisory position to demonstrate many level 3 competences, but they wouldn’t be in such a position and cannot secure such a position, if they are not yet considered competent. Similar issues pertain to client safety in Social Care, even at lower levels of the awards. Care workers will not (should not?) be in a position to exhibit evidence of safe practice until they are already competently safe.

Policy-related questions arise about whether the goal of defining current national standards is still appropriate to workplace activities (and if not how they can be updated quickly) and, in turn, whether simulation is acceptable and if so, to what extent. For example, changing a clutch in a college workshop because no such job has occurred in the workplace would seem to be acceptable (if still deemed necessary); ‘pretending’ to do a client fitness appraisal on a colleague that one works with every day is perhaps less appropriate.

References

- Torrance, H., et al (forthcoming) *The impact of post-16 assessment systems on achievement* (London, LSDA)
- Wolf, A. (1995) *Competence-based assessment* (Buckingham: Open University Press)

4. Scotland: Assessment 3 to 14

Carolyn Hutchinson

Context

Ambitious, Excellent Schools, published in November 2004, provided a broad policy framework for transforming education in Scotland's schools, building on the themes already established in the National Priorities (2000) and *A Partnership for a Better Scotland* (2003). The framework aspires to ensure that all young people fulfil their potential at school, and gives a clear commitment to putting the learner at the centre of education. It is recognised that assessment has a crucial part to play in achieving this outcome, because it can give learners, and those who teach and nurture them, the feedback they need to improve their learning.

Assessment is for Learning ('AifL') began in early 2002 following a review of assessment 3-14 by HM Inspectors of Schools and consultation on the report's recommendations. Published in December 2000, the report on consultation set out views of respondents about the changes that were needed in the system of assessment for the primary and early secondary years in Scotland:

- change should be introduced as 'evolution, not revolution', building on current strengths in practice; it should be manageable, properly resourced and supported;
- classroom assessment to support learning, rather than measurement of learning for monitoring and accountability purposes, should have much more emphasis;
- there should be a common national framework for record-keeping and reporting, although schools wanted the discretion to adapt the framework to their own circumstances;
- improvements should be made to arrangements for National Testing;
- in introducing changes to the system of assessment, due attention should be paid to recent research evidence.

An Assessment Action Group, chaired by the Deputy Minister for Education, was set up in January 2002 with the overall aim to "provide a streamlined and coherent system of assessment to ensure pupils, parents, teachers and other professionals have feedback they need about pupils' learning and development needs." To develop the new system, the programme set out to:

- develop good professional practice and confidence in assessment amongst teachers so that their judgments would be dependable;
- put in place credible quality assurance of teachers' judgments locally and nationally, as part of understanding and sharing standards;
- monitor national attainment in a way that provided accurate information about overall standards and trends and at the same time promoted good classroom practice.

Ten linked projects were initially established, each focusing on one of these three aspects of assessment. Groups of teachers were given small grants to undertake classroom-based action research projects to develop their own understanding about assessment practice, skills in integrating assessment into classroom practice to support

learners, and some skill in self- and peer-evaluation. The outcomes of these initial projects, together with feedback from formal evaluations and consultation, were used to review AifL and to bring the various aspects of assessment investigated back together into a streamlined and coherent system. Assessment for learning and assessment for monitoring and accountability would be complementary, rather than in opposition.

Design features

AifL has sought to 'join up' research, policy and practice - gathering evidence from research and monitoring activity; using the evidence to develop informed policy; and working with and supporting practitioners and schools to build informed communities of practice. There were three significant influences on the design of AifL: reflections on implementation of previous Assessment 5-14 policy initiatives (1990); research about assessment for learning in Black & Wiliam's *Inside the Black Box* (1998); and work on transformational learning, in particular Senge and Scharmer's analysis of community action research approaches (2001).

Research in assessment suggests that learners learn best, and attainment improves, when learners:

- understand clearly what they are trying to learn, and what is expected of them;
- are given feedback about the quality of their work, and what they can do to make it better;
- are given advice about how to go about making improvements;
- are fully involved in deciding what needs to be done next, and who can give them help if they need it.

These ideas underpin the three strands of work which contribute to becoming an 'AifL school'.

In the AifL programme learning is also about transforming communities of practice and ownership by teachers is seen as being key to promoting and sustaining change. Support has been available to schools and local authorities through a number of sources. These include Scottish Executive Education Department funding for local authority co-ordinators and development officers; national development officers and staff from the Scottish Executive, Learning and Teaching Scotland and the Scottish Qualifications Authority working in partnership; assessment focused consultancy support, also through Learning and Teaching Scotland. In addition, interested staff from Scottish Faculties of Education have worked as a network to support teachers and schools involved in the programme to develop 'action research' approaches, to access recent research about the area they were investigating, and to reflect on the impact of the AifL programme on their own classroom practice.

Progress to date

195 schools were involved in the initial phase of AifL. By December 2004, local authority reports on the number of schools involved in AifL through associated schools groups (secondary schools and their associated primary schools and early years establishments) had increased this number to 1,581 schools. Working in ASGs has emphasised the importance of professionals working together and building communities of practice.

The outcomes of projects are captured in evaluative case studies which are available in the Assessment Online Toolkit, a resource aimed primarily at Scottish classroom teachers and school managers, but which will also be of interest to local authorities, researchers, trainee teachers, parents and pupils.

For schools and teachers, three main strands of internal assessment activity now underpin the programme as it is introduced across Scotland: assessment FOR learning (formative assessment), assessment AS learning (personal learning planning) and assessment OF learning (understanding and sharing standards). Each strand has a number of 'key features' attached to it and these are detailed in support materials for the 'AifL school – a place where everyone is learning together'. There are ten key features of an AifL school:

Assessment for learning:

- Our classroom assessment involves high quality interactions, based on thoughtful questions, careful listening and reflective responses.
- Our pupils, staff and parents are clear about what is to be learned and what success would be like.
- Our pupils and staff are given timely feedback about the quality of their work and how to make it better.
- Our pupils and staff are fully involved in deciding next steps in their learning and identifying who can help.

Assessment as learning:

- Our pupils and staff practise self- and peer-assessment.
- Our pupils and staff help to set their own learning goals.
- Our pupils and staff identify and reflect on their own evidence of learning.

Assessment of learning:

- Staff use a range of evidence from day-to-day activities to check on pupils' progress.
- Staff talk and work together to share standards in and across schools.
- Staff use assessment information to monitor their establishment's provision and progress, and to plan for improvement.

External national monitoring is now carried out by means of a sample survey, the Scottish Survey of Achievement, rather than blanket national testing, so that accountability no longer directly drives classroom activity. The first SSA, a survey of English language and core skills, was carried out by the Scottish Executive in May 2005. Further surveys will monitor attainment in social subjects enquiry skills (2006), science (2007) and mathematics (2008), and core skills in each of those subject contexts.

At the same time, work is continuing to develop and extend an on-line national bank of assessments, using the assessment materials from the SSA. The assessments are intended for internal use by teachers as additional evidence to confirm their own considered judgments about pupils' progress through the current 5-14 levels of attainment A-F. The domain-referenced assessments and tests in the bank are therefore intended as a means for schools to quality assure teachers' judgments, rather than as definitive measures of attainment. Schools will be able to benchmark the results for their own pupils against the results for the survey as a whole, and use them as feedback to inform planning of future programmes of work in the subject areas concerned.

Proposals for development

Scottish Ministers confirmed their commitment to introducing AifL into all Scottish schools by 2007 in *Ambitious, Excellent Schools*, their vision for Scottish education. Further detail for schools and authorities about the new arrangements and about the roles of local authorities, schools and early years centre managers in assessment was outlined in *Circular No.02* (June 2005).

Information and resources are being provided through AifL to help schools and local authorities to develop their assessment policy and professional practice in assessment for, as and of learning, including the Assessment Online Toolkit and the on-line bank of assessment tasks. A self-evaluation toolkit, linked to the HMIE publication *How Good Is Our School* and associated quality indicators, will be piloted in schools during session 2005-06.

AifL is gradually changing the perception of what can be done in classrooms and the impact of changing classroom practice on young people. The programme seeks to be responsive to ongoing feedback and evaluation received from case-studies, local authority network meetings, consultation events for parents, for pupils and for staff, national consultation, and formal, independent evaluations. Issues identified through these sources have been used to inform the next steps for AifL, reflecting the philosophy behind AifL itself which gave considerable freedom to schools and teachers to develop practice within their own context at a pace and in a manner that suited local needs. The AifL team will continue to work with authorities and school managers in the creation of a single, coherent assessment system to promote assessment for learning and to provide assessment information for monitoring and measurement. Support will continue to be provided through a variety of mechanisms including continued funding, AifL newsletters, national and regional events, the 'AifL school' resource pack, and collaborative support from LTS Development Officers and SEED officials.

Feedback from internal evaluation suggests that key features of successful projects within the programme to date have been:

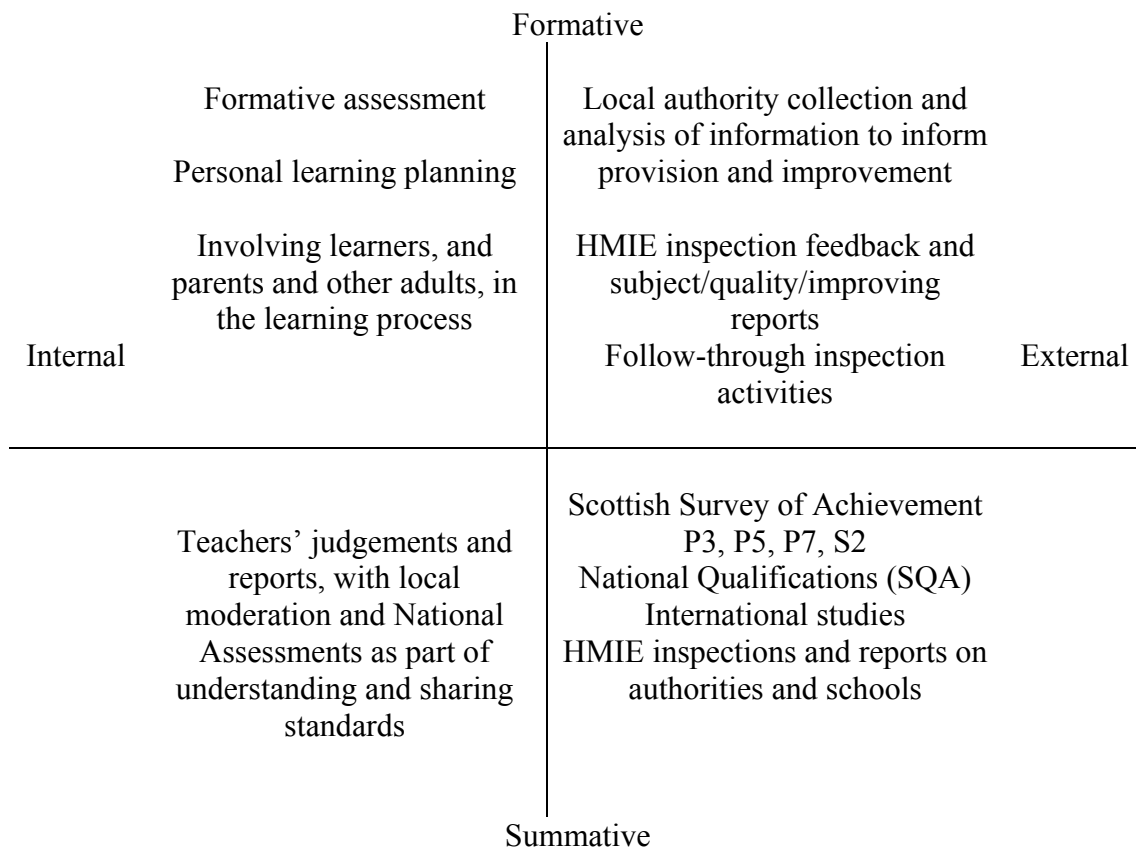
- Educational integrity: focus on *learning*; clear effects on pupils' self-esteem and classwork; teachers' increased confidence in their teaching and in formative assessment approaches.

- Collaboration within and across communities involved in the development: being part of a development team and a network where you can discuss what you have tried out; practical guidance and modelling from other teachers; effective management of the conditions to promote assessment for learning, including committed leadership; effective resourcing and organisation of time.
- Reassurance in respect of worries about formative assessment for learning not being regarded as valuable “accountability” activity: development of agreed departmental or stage approaches to it; clear valuing of assessment for learning throughout the education system – by the school, the local authority, national policy and the research community; the stimulating effect on their thinking of contributions from distinguished researchers.
- Real involvement: the impact of the experience of “credible” teachers who had already used formative assessment; autonomy in trying out, discussing and adapting approaches; thus, realisation by teachers that assessment for learning builds on their own familiar professionalism.

There are four main continuing challenges to take into account as the AifL programme is developed more widely across Scottish education authorities and schools.

- Mere planning for development is inadequate without action to promote effective assessment.
- The growth process is inevitably complex and should not be over-simplified or "rationalised".
- The collaborative interaction of the practice, management, policy and research communities should continue, without any one community's distinctive perception of its role, or of the needs of the programme, dominating.
- In-depth understanding of the principles of assessment for learning, rather than simply use of "strategies" or "techniques", is crucial for all of the communities involved.

The national assessment system in Scotland



5. Wales: end-of-stage assessment at Key Stage 2

Richard Daugherty

Context

The Welsh Assembly Government is currently introducing new statutory arrangements for assessment at Key Stages 2 and 3 to take effect over a period of three years. This case study focuses on the reporting of pupils' overall performance in the core subjects of English, Welsh, Mathematics and Science as judged by their teachers at the end of Year 6, the last of the four school years that comprise Key Stage 2. Until 2004 this was done in two distinct ways, with a pupil's 'levels of attainment' being based both on the results of statutory tests and on the 'best fit' judgments of teachers. From 2005 Key Stage 2 tests are being phased out.

The end of Year 6 not only marks the end of a National Curriculum key stage. It is also the point at which all pupils in state schools in Wales transfer, at approximately age 11, from primary schools to non-selective comprehensive schools (there are no selective secondary schools and no middle schools). In line with a provision of the 2002 Education Act, the Welsh Assembly Government is to require all schools to have transition plans in place from 2008.

Another aspect of the set of changes to statutory assessment recommended by the Daugherty Assessment Review Group in 2004 was 'skills profiling'. Skills profiles, supported by tests, would begin in Year 5 and, suitably modified, then be carried through the pupil's transition to secondary school.

The changes to the assessment and reporting of pupil attainment in the core subjects that are discussed here have been developed alongside plans for skills profiles and in anticipation of transition planning.

Design features

In the system that was in place until 2004 teachers did judge which of the National Curriculum 'levels of attainment' best fitted the standard in each attainment target in each core subject attained by each of their pupils towards the end of Year 6. However, those judgments were not supported by formal procedures to ensure that such judgments, arrived at separately by each teacher, were consistent. Moreover, the reports of teacher assessments were overshadowed by the results of tests in the same subjects. In the new system as in the old, a child's levels of attainment, supplemented by other information, will be reported to (a) her/his parents and (b) the secondary school to which s/he is to transfer.

In order to achieve an acceptable level of consistency in teachers' judgments the Review Group recommended that schools should be grouped on the basis of notional secondary catchment areas with each primary school linked for the purpose of these moderation procedures to a particular secondary school. Primary and secondary teachers from each group of schools would meet twice in each school year for agreement trials using pupils' work in the subjects being assessed. Guidance material

and procedural advice would be published by the national agency, ACCAC, which would also monitor the effectiveness of the procedures. An ‘acceptable level of consistency’ in this context is interpreted here as being such as to give the secondary schools sufficient confidence in the levels of attainment reported to them for the schools to make use of them as benchmark indicators for subsequent progress. Local authorities would be encouraged, but not required, to facilitate cross-catchment and cross-authority arrangements with a view to maximising the convergence of teachers’ judgments on a wider basis.

Though the use of aggregate pupil attainment data as an indicator of the quality of schools has been contentious in Wales, as in England, it has not been the practice of successive governments in Wales to publish such data on individual primary schools. The extent to which attainment data generated by teachers’ judgments at the end of Key Stage 2 is used in future as a school performance indicator will be a significant factor in determining whether this component of the new system is able to fulfil its primary purpose, i.e. to report reliably on the attainments of individual pupils. The use of aggregate data for monitoring and evaluation will need to be restrained if the negative ‘backwash’ effects of such use on learning and teaching are to be minimised.

The design of the new end-of-stage arrangements for judging attainment has recognised that both the commitment and the expertise of the teachers involved will be crucial to its success. In advance of full implementation of the new arrangements in 2007 all the teachers involved will need opportunities for relevant training and continuing professional development (CPD). Ongoing CPD thereafter will also be required to support teachers in working through issues that arise in the course of implementation. Subject-based guidance developed by ACCAC will be published to help teachers make whole subject judgments together with procedural guidance to explain the use of agreement trialling.

Properties

Validity

Questions of construct validity in this component of the system are the same as the questions that have been grappled with ever since the introduction of National Curriculum assessment. What is new about these arrangements is that the focus has shifted from the dubious construct validity of test-based whole subject judgments (less of a problem in relation to Welsh, because of the way that subject has been assessed, than in the other three subjects) to the hitherto undervalued teacher-based judgments. What is also new is that moderation procedures will be in place in every school to support teachers in developing a shared sense of the constructs that are implicit in the National Curriculum levels of attainment.

Issues of consequential validity were at the heart of the arguments for changes to the system that was in place until 2004. The decision to discontinue the former test-based (‘SAT’) judgements of whole subject performance was much influenced by reports of the negative consequences of those tests for learning and teaching in primary schools, especially in Year 6. In the planning of the new system the primary purpose of Year 6 subject judgments has been stated in terms of them making a valued contribution to informing secondary schools about the attainments of their pupil intake. This,

alongside consideration of other consequences (some, inevitably, unintended), provides a clear basis for evaluating the new provisions. A discussion of the extent to which the new arrangements are satisfactory in consequential validity terms should be in terms of both (a) a reduction in the negative effects attributable to end-of-stage tests and (b) an improvement in the extent to which Year 6 subject judgments have a valued role in pupil transfer.

Reliability

A comparison of the new arrangements with those they have replaced would include well-founded scepticism about the reliability of the SATs (see case study 2 for a discussion of this). However, the key reliability consideration for the new system of teacher-based summative assessment will be whether the group moderation procedures, based on secondary catchments, have the effect of building the confidence of secondary schools in the reliability of the judgments made by teachers in primary schools. That limited aspiration of an ‘acceptable’ level of consistency in teacher judgments is the declared goal but comparability of teacher judgments across different catchments will no doubt continue to be an issue when the new system is debated.

Practicability

Successful introduction of all the new assessment arrangements in Wales will depend on each aspect being well planned, resourced and implemented. One dimension to their practicability will be whether the system’s expectations of teachers are seen by them as making unreasonable demands on their time. Implementation of a new assessment system which, crucially, also give a high priority to the teacher’s role in using assessment directly to support each pupil’s learning (‘assessment for learning’) comes at a time when teacher workload is a high profile political issue.

Issues arising in implementation

It is already clear that several issues will have to be addressed if the new arrangements are to be implemented successfully. These include:

- *Retaining a focus on the individual.* The use of summative teacher assessment will need to remain focused on the primary purpose of this component of the system as envisaged by those who designed it, viz. the use of whole subject attainment data about individual pupils as part of the profile that each pupil carries with him/her on transfer to secondary school. The moderation procedures that have been put in place have been designed with that purpose in mind.
- *Minimising the impact of monitoring and evaluation.* If data on the whole subject attainment of individuals were to continue to be seen by many of those involved mainly as a source of aggregate indicators of system performance, the rationale for the intended approach to end-of-stage assessment will be undermined.
- *Prioritising formative assessment.* The summative components of the overall assessment arrangements will need to be neither presented nor resourced as if they were of greater importance than the development of ‘assessment for learning’, is a key recommendation of both the Daugherty Assessment Reform Group and ACCAC.

- *Teacher ‘ownership’ of the system.* The national framework for this component of the assessment system will need to be flexible enough for the teachers involved to feel that they are actively engaged in shaping the system as well as making specific judgments about their pupils’ attainments. There is clear evidence from research (see ASF Working Paper 2) that teacher ‘ownership’ of their part in summative assessment is essential for successful implementation.
- *Clear but flexible guidance.* National guidance on subject judgments and on procedures will need to be clear enough to reinforce the message that this is a commonality of approach across a national system without becoming unduly prescriptive.
- *Between catchment moderation.* Effective cross-catchment and cross-local authority arrangements will need to be sufficiently widespread and showing evidence of influencing outcomes for the new procedures to be seen and accepted as a national system but one that is locally administered.
- *Evaluation* procedures, including self-evaluation within catchments, will need to be in place to feed back into system improvement.
- *Resources* will be needed to cover the staffing and other costs incurred in moderation, infrastructure development and maintenance and evaluation.

What lies ahead for all the new assessment arrangements in Wales, of which this is only one element, can be summarised in terms of the various types of challenge that will have to be faced – technical, professional and political. There are technical challenges such as the design of a system for maximising consistency in teachers’ assessments that is fit for purpose. There are professional challenges for teachers in developing their skills of assessment and in taking ownership of the procedures associated with end-of-stage assessment. And there are political challenges for those responsible for the education system in Wales locally and nationally, officials and elected representatives. These include putting summative assessment in context alongside giving greater priority to assessment for learning, resourcing the teacher development that is needed and reining in the tendency to see pupil assessment mainly in terms of monitoring system performance.

10. England, Wales and Northern Ireland: GCSE

Gordon Stobart

Context

The General Certificate of Secondary Education (GCSE) is the examination taken by 16 year olds at the end of compulsory schooling in these three countries of the UK. It was introduced in 1988 to replace the two-tier system of GCE O levels and the Certificate of Secondary Education (CSE) and these origins are reflected in the grading system. GCSE officially has A*-G pass grades although in practice only A*-C (the O level equivalent grades) are treated as a pass. It is high-stakes for the students, as progression and selection is largely based on it, and for their schools which are evaluated in relation to the grades their students are awarded. Five grades A*-C in this subject specific examination is the key benchmark in the evaluation of school and system performance, with students typically taking 8-10 GCSEs. As these are largely externally marked, the examining system is put under extreme pressure each summer, when national tests and GCSE A levels are also sat and marked.

Design features

Each GCSE involves a detailed subject specification which, where appropriate, is based on the national curriculum key stage 4 requirements. Both the core content and the schemes of assessment are controlled by the regulatory authorities¹. The five awarding bodies² can offer only a limited and prescribed number of specifications in each subject (to meet comparability concerns). As a consequence the GCSEs offered in a particular subject are similar in both content and assessment.

The assessment scheme in most subjects involves 'tiering' which is intended to help differentiate across a wide range of attainment and make the examination a positive experience for low attaining students. The standard tiering model involves a student entering either the foundation tier (grades C-G) or the higher tier (grades A*- E). There is often common coursework across both tiers, though there is usually a ceiling on the grade that can be achieved for foundation tier students.

A typical GCSE scheme of assessment involves the student taking two 1-1.5 hour examinations, which are externally marked, and submitting coursework which is teacher assessed and externally moderated. The examinations are generally open-ended and will involve short and structured written responses alongside some longer written answers. The examinations typically account for around 75% of the final mark (less in more applied subjects) with the coursework contributing the rest. Coursework ranges from specific practical activities to written assignments or portfolios. In order to increase reliability many coursework task are designed by the awarding body and are completed by students in the classroom within a specified time period.

Properties

¹ QCA in England, ACCAC in Wales and CCEA in Northern Ireland

² Edexcel, OCR and AQA in England, WJEC in Wales and CCEA in Northern Ireland

Validity

At one level it could be claimed that there is high construct validity because the tests faithfully mirror the specifications – which are designed with the test in mind. At a more general level we can ask how these specifications sample the broader subject domain. Judgements here will vary from subject to subject. For example, there are long-standing debates about the role of literature in English and about the content of the history curriculum. At the level of aims and values there is an argument³ that the curriculum on which the specifications are based often does not meet its own aims, for example encouraging collaborative skills.

Coursework offers, in principle, an opportunity for a fuller sampling of the domain and the subject skills involved. However, successive moves to contain and standardise this element have meant that in practice it is often now little more than a prescribed classroom exercise which has limited validity and fitness-for-purpose.

Reliability

GCSE examinations are subject to the same critique of their reliability as other tests but no information on their reliability is publicly available. There are extensive procedures in the regulatory bodies' Code of Practice which range from paper setting through to awarding processes and appeals. There are also Code of Practice procedures for coursework which include moderation arrangements.

However, the examination papers are not pre-tested and marker reliability remains a concern in many subjects. In order to make the process more transparent the marked scripts are now made available to schools.

The reliability of teacher assessed coursework is currently a key concern (see below). While at earlier stages the focus was the reliability of the teachers' marking, this has given way to concerns about the authenticity of the students' work, given the ease with which it can be downloaded from the internet.

Practicability

There is wide recognition, including from the chief executive of the Qualifications and Curriculum Authority in England (QCA), that the current assessment regime is unsustainable. GCSE contributes a large proportion of the 20 million plus scripts that have to be marked by examiners each summer. While attempts are being to develop e-marking this only tackles some of the transport logistics.

Other practicability issues for schools are costs, resources (eg loss of halls and gyms during the six week examination period) and the number of student teaching days lost through examinations and preparation for them.

Impact

The primary purpose of GCSE is to certificate individual students at the end of compulsory schooling. It largely succeeded in combining into a single examination

³ See J.Wilson (Ed.) Rethinking Curriculum 2003

the prestigious GCE O Level⁴, taken by a minority of students, and the CSE⁵ which was taken by most. A practical consequence was that grades A-C, considered equivalent to an O Level pass, have become the passing grades, even though grades D-G are technically passes. As none of the national systems has any form of school graduation or certification, GCSEs are the key indicator of individual achievement at the end of compulsory schooling and is used as the basis for further progression in education..

Accountability and the decline of coursework

Introduced in the late 1980s, GCSE incorporated some of the more innovative elements of CSE, in particular a coursework component in almost all subjects (its introduction was delayed for GCSE mathematics). The most popular GCSE English syllabus during this early period was 100 per cent teacher assessed. The use of coursework assessment had been strongly backed by the Secretary of State for Education who oversaw the introduction of GCSE, Sir Keith Joseph. But, by 1991, the intervention of the then Prime Minister, John Major, led to the scaling back of coursework. This can be seen as a move to limit the contribution of teachers to GCSE grades which would increasingly be used for school accountability purposes.

The development of performance indicators using aggregate data on the GCSE performance of pupil cohorts means that schools have become increasingly focused on meeting targets and improving their position in the annual performance tables. This has had a considerable impact on schools (a consequential validity issue) as they seek to optimise their results.

With only half the cohort gaining five or more grade A*-Cs and only around a third achieve five 'passes' which include English and mathematics schools have sought alternative ways to boost their scores. One popular one has been to enter for a vocational qualification (GNVQ Intermediate⁶) which is regarded for accountability purposes as equivalent to four grade C GCSEs and involves a much higher proportion of teacher assessed coursework. Students then only need one other GCSE at grade C to reach the target. The highest performing comprehensive school in England has pioneered one such course in Information Technology and many of the 'most improved' schools in 2005 had taken GNVQs. This unintended consequence has recently led the government in England to make GCSE English and mathematics a compulsory part of the 'five A-Cs'.

Monitoring of standards

The monitoring role of GCSE results is more ambiguous. While improved results are generally taken as showing improved standards, there is an underlying concern that this may be a consequence of reduced examination demands. With that in mind any increase from one year to the next of more than 2 per cent in the higher grades is likely to be investigated by the regulatory agencies.

⁴ The General Certificate of Education Ordinary level which was designed for the top 20-30 % of the 16 year old cohort and which rarely included coursework.

⁵ Certificate of Secondary Education, a more innovative and often locally developed examination intended for less 'academic' students which had a strong coursework component.

⁶ General National Vocational Qualification, a full-time course originally intended for students in Further Education colleges.

The crisis of confidence in teacher assessed coursework

While the earlier concerns about coursework were essentially about the reliability of teacher assessment, especially in an increasingly target-driven culture, the more recent concerns have been more about the risks of plagiarism and collusion. At the centre of this is the ease with which ‘exemplar’ essays can be downloaded from the internet. Because the assignments topics are now explicitly specified, prepared answers become easier. There is anecdotal evidence of there being around 5000 model answers to the mathematics coursework tasks on the internet.

An investigation by the Qualifications and Curriculum Authority (QCA, 2005⁷) led the Education Secretary to call for an urgent review of coursework. The QCA findings found that internet copying ‘cannot be controlled’. There was also concern about the extent to which teachers helped through highly detailed writing frames, templates and check-lists. This has led to ‘coursework cloning’. In addition, QCA’s survey found that 63 per cent of parents helped in some way with coursework, with 39 per cent helping to find information and 26 per cent supervising the work.

The likely consequence of this is that coursework is likely to be further restricted in GCSE examinations, possibly by insisting on more controlled conditions. These may further weaken the validity and fitness-for-purpose of coursework since it may become little more than a long examination. It may even disappear from mathematics and science.

What the crisis in GCSE coursework illustrates is the importance of public credibility in high stakes teacher assessment. Attempts to address this by narrowing the scope of coursework could further undermine its validity. The increased standardising of topics may also have, unintentionally, weakened coursework reliability as plagiarism and collusion become easier.

⁷ Report can be found on www.qca.org.uk/15525.html