EXPERIENCES OF SUMMATIVE TEACHER ASSESSMENT IN THE UK

A review conducted for the
Qualifications and Curriculum Authority

John Wilmut

Commissioned by the Qualifications and Curriculum Authority.
QCA, 83 Piccadilly, London, W1J 8QA. United Kingdom.

Contents

Foreword

This review has been undertaken for the Qualifications and Curriculum Authority (QCA) under Purchase Order 35502 of 4 November 2003. It seeks to explore experiences of summative teacher assessment in the UK with the specific aims of

- mapping out a full range of teacher assessment models which have been implemented, on a significant scale, in the UK since 1950, indicating their particular operational characteristics and highlighting their key dimensions of difference (e.g., task-based vs. observation-based, high stakes vs. low stakes, holistic vs. atomistic judgements, record-intensive vs. record-un-intensive, master/non-master judgements vs. graded scale, etc.);
- providing key illustrative instances of implementation, identifying the aspirations which were held for each of these teacher assessment models (e.g., why they were implemented, and how they were expected to function), and the extent to which those aspirations were realised in practice;
- exploring the reasons why these examples were successful or unsuccessful in meeting aspirations in their particular (social, cultural, political, educational) contexts, highlighting the features of those contexts that had a particular impact upon their success or failure (e.g., demands for reliability/comparability/validity, financial constraints, time constraints, training and moderation requirements, scale of operation, positions adopted by influential players);
- more generally, identifying, for each of the teacher assessment models, any (social, cultural, political, educational) contexts for which they seem particularly suited or unsuited;
- speculating upon possible future directions for assessment 14-19, to identify key messages for policy makers.

The work is set against the background of a statement from the Working Group on 14-19 Reform that seeks to promote the increased use of teacher assessment in, for example,

> 37. We are particularly keen to reinforce the role of assessment which is based upon the professional judgement of teacher and trainers. We do not believe that, with effective training and monitoring, internal assessment is necessarily less reliable than external assessment. More effective internal assessment need not be accompanied by an increase in workload for teaching staff. For instance, we believe that better use could be made of assessment which is already undertaken by many teachers and trainers of work done by learners as a natural part of their course, rather than relying on a greater volume of externally-prescribed coursework tasks. These changes 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 to re-establish the credibility of such assessment as a reliable tool for judging the achievement of young people.

In following this brief I have placed some artificial constraints on the work, following advice and help from staff at QCA. For example, this review does not include an exhaustive literature review, so that I have been selective in the sources that I have accessed, with a view to illustrating general issues of principle rather than discussing every aspect of all teacher assessment activity since 1950 (if that were even possible!). Then, the emphasis here is on teacher assessment for summative purposes; it would be wrong not to examine the

relationship between this type of activity and formative assessment (or assessment for learning) but I have not discussed many issues associated with this use of assessment.

Encouraged by QCA colleagues I have strayed well outside the 14-19 curriculum where (particularly in relation to national curriculum assessment) there are important aspects of teacher assessment to consider. However, I have tried to focus those ideas on the issues concerned with curriculum and assessment 14-19, particularly in relation to qualifications in that sector. Similarly, I have incorporated ideas and evidence from vocational and occupational qualifications, even though some of these operate beyond the 14-19 sector, chiefly because they offer a different sort of experience, particularly in defining the role of the 'teacher', in the types of evidence of attainment that are generated, and in the methods of quality assurance that operate. I strayed only a little outside UK experience, mainly because another parallel piece of work is looking at what the UK may learn from experience of teacher assessment in other countries.

This is a short piece of work which could not have been accomplished without help from a number of people. I am particularly grateful to Paul Newton and colleagues at QCA for information and advice. I have been greatly helped in discussions with Wynne Harlen and my understanding was considerably enhanced by my involvement in the Nuffield/Assessment Reform Group seminar, held on 12-13 January 2004 in Cambridge as part of the *Assessment Systems for the Future* project that she is directing, and which will run until December 2005.

The views that I express here are mine and may not be held by others that I have consulted, quoted or listened to.

June 2004

John Wilmut
The Old Post Office
Bray Shop
Callington
Cornwall PL17 8PZ, UK
tel +44 (0)1579 370736
fax +44 (0)1579 371064
email: wilmut@aol.com

Executive Summary

Each of the main discussion sections of this report (2, 5-11) ends with a summary. The taxonomy (Section 3) and examples (Section 4) have not been summarised and Section 12 is then a discussion that attempts to address four questions:

- What general principles, grounded in appropriate frameworks of assessment, learning and progression, might underpin models for teacher assessment for summative purposes?
- What are the threats and risks (in relation to reliability, validity, public confidence, etc.) associated with moving towards a far greater reliance upon teacher assessment in general qualifications?
- How feasible is it for teacher assessment to support both formative and summative purposes?
- How, if at all, does teacher assessment address the issue of the overall burden of assessment on teachers, students, awarding bodies, etc.?

The section summaries have been consolidated into this Executive Summary and should be read in conjunction with Section 12.

The development of internal assessment

In relation to the development of public examinations:

- Examinations and tests have come to be regarded as objective, fair, equal for all candidates, and as having high reliability. They have been managed and delivered very successfully and have achieved considerable status that has been strengthened by the beliefs that they provide motivation for learners to work and that they ensure that educational standards are maintained. (para 2.4)
- There has been no planned development of internal assessment over the last 50 years – it has been a component of a number of initiatives but has rarely, if ever, been viewed as a whole. Until recently there has been almost no perspective on internal assessment that has embraced all general, vocational and occupational qualifications. (para 2.6)
- There is still only a limited understanding of the relationships between internal assessment for summative purposes within qualifications, internal assessment for summative purposes within the national curriculum and assessment for learning. (para 2.6)
- There are risks in attempting to treat internal assessment as an integrated phenomenon and in generalising from one context to another. (para 2.6)
- Traditionally it has been argued that the validity of external school examinations would be enhanced if some of the objectives were to be assessed by class teachers (para 2.7) and that such a move would bring summative assessment closer to the formative assessment of the classroom, enhancing the latter. (para 2.8). It has also been suggested that internal assessment for summative purposes involves teachers in taking responsibility for assessment of their pupils' performance. (para 2.14)

In relation to vocational and occupational qualifications:

- Vocational and occupational qualifications have involved the collection of evidence of competence that is related to highly detailed and fragmentary performance statements. This led to a very complex, cumbersome and bureaucratic assessment system, dominated by paperwork. (para 2.23)
- There has often been a belief that making the statements which specify competence more specific and more exact will result in greater clarity and more valid and reliable assessment. This has not been the case: re-writing has resulted in more complexity and less clarity, as well as further obscuring the overall concept of competence. (para 2.24)
- Concepts of a 'standard' derive from a complex and ad hoc mix of experience, liaison with colleagues and others, and some informal sharing of assessment judgements. (para 2.24)
- Assessment activities have turned into strong compliance with the assessment criteria. This was a pragmatically rational response to the conflicting pressures they were under that was promoted as an educational goal in its own right. (para 2.25)

In relation to authentic assessment:

- Authenticity is not achieved by just injecting a bit of internal assessment into an external examination. (para 2.29)
- If the primary goal is to maximise reliability then internal assessment might be an inappropriate tool. If the primary goal is to harness a powerful tool for learning then internal assessment may be essential. (para 2.30)
- There is a significant difference between teachers' use of external assessments over which they have no control and their use of external assessments that they choose and which they can entirely control. (para 2.31)

A rationale for internal assessment

In relation to the purposes underlying teacher assessment for summative purposes:

- A range of justifications has been offered including that it enables objectives to be assessed that are inaccessible to written examinations; it brings teacher assessment closer to classroom assessment practice; negative backwash on the curriculum is reduced; a closer link between learning and summative assessment is established; the values and pedagogic requirements of teachers are better met if they have some control over the summative assessment; partnerships between teachers and examination boards are beneficial; teachers' assessment expertise is enhanced. (para 5.1)
- It is said that internal assessment is the only appropriate way to determine learner competence in workplace and other vocationally-related settings. (para 5.1)
- The primary issue in discussing purpose is policy and control: where should the decision-making and responsibility lie for the conduct and outcomes of the teacher assessment? Discussions about other issues (such as about the relationship between assessment for formative and summative purposes, about validity and reliability and about quality control) are incapable of resolution unless there is some clarity in principle about this. (paras 5.2 & 5.7)

- The purposes for internal assessment are not static, and not always shared between teachers and the awarding body. However, involving teachers in the development of shared common purposes for coursework is difficult and costly in large-scale public examinations. (para 5.7)
- Perceptions of the purposes of coursework need continual re-defining for various stakeholder groups. (para 5.7)
- The values made explicit in a syllabus are not necessarily those that underpin learning in the classroom and the formative assessment used there. (para 5.7)

In relation to the engagement of teachers in summative assessment:

- In the past many teachers were engaged with summative assessment for curriculum reasons rather than primarily as a process of assessment reform. In some situations teacher assessment was teacher-driven and articulated in relation to specific pedagogic demands. (para 5.5)
- Where teachers have become involved in the use of examination coursework that has been imposed from outside their attitudes have been less clear and often less supportive. Rationales that connect external control over coursework with better control over the validity and reliability of the assessment have carried less weight and many teachers have quickly become concerned about the additional workloads involved and about the technical details of the conduct of the assessment processes. (para 5.6)
- Training and support provided by awarding bodies and others has been less concerned with developing teachers' generic assessment skills than with the mechanics of the conduct of a specific qualification and its quality control. This has sprung from a concern to achieve the highest possible reliability coupled with the minimum demands on teacher time. (para 5.6)
- Balances need to be struck between teachers and others stakeholders in relation to the degree of control over the choice and specification of tasks, acceptable levels of validity and reliability and the relationship between internal and external control of quality. There do not seem to have ever been rational and systematic debates of these factors and it is hard to see how progress can be made without it. (para 5.10)

In relation to getting a better understanding:

- Reliability and bias in internal assessment should be more accurately placed in the context of the reliability and bias in all forms of assessment. The reliability and bias associated with different approaches to internal assessment should be studied more exhaustively so that appropriate choices of method can be made or patterns of support developed that will effectively reduce the problems. (para 5.11)
- There is a need to explore whether mechanisms for creating summative assessments from some set of formative assessments made by teachers will overcome problems of workload and problems of validity and reliability at levels required for summative reporting purposes. (para 5.13)
- Research is needed on whether there is insufficient replicability in culling teachers' existing assessments across many domains to generate the required levels of reliability for summative purposes and whether there is bias in the assessments that cannot be routinely detected or controlled. (para 5.13)

- There should be further exploration of the proposition that external summative assessments narrow the curriculum. (para 5.14)

Internal assessment for formative and summative purposes

In relation to the tension between formative and summative purposes:

- For many years there has been seen to be a tension between the purposes of formative and summative assessment, with the view that it is better to keep them apart. (para 6.3)
- There have been said to be tensions in the role of the teacher when involved in both formative and summative assessment. (para 6.3)
- More recently, teachers have drawn formative and summative purposes together by culling test materials that have been designed for summative purposes for monitoring and formative uses. It has been argued that formative and summative roles must be combined in order to escape the dominance of external summative testing. (para 6.4)
- Where students need to generate evidence during a programme and have it assessed teachers tend conflate formative and summative functions of internal assessment. (para 6.4)
- It has been said that continuous assessment cannot function formatively when each attempt or piece of work submitted is scored and the scores are added together at the end of the course. This practice tends to produce in students the mindset that, if a piece of work does not contribute towards the total, it is not worth doing. (para 6.4)
- We cannot say that any test or instrument sourced externally will automatically be summative and that any piece of less formal assessment done in a classroom will automatically be formative or for learning. Nor can we assume that an instrument developed for external summative purposes and used uncritically or inexpertly will satisfactorily perform a useful formative role. (para 6.5)
- We cannot have good formative assessment in situations where teachers' efforts are not valued, respected and trusted. (para 6.6)
- There is a need to get formative processes right first since it is a prerequisite for both good learning and good summative assessment. (para 6.6)
- A resolution of the tension between formative and summative assessment will require teacher expertise in assessment that is concerned with principle and purpose and not just with process. (para 6.5)
- We should not under-estimate the levels of professional expertise in assessment required of teachers if they are to become key players in the resolution of the tension between formative and summative purposes for assessment rather than just implementing externally controlled initiatives. (para 6.7)

In relation to the conduct of formative and summative assessment:

- Teachers find it difficult to break out of a mode of assessment that is concerned with making summative judgements. (para 6.8)
- Teachers' understanding of what counts as valid evidence becomes the framework for the knowledge that students have transmitted to them. Students do not tell

teachers what they actually know, but what is seen to fit into the frameworks as they understand them. (para 6.8)

- Formative assessment processes may be convergent (such as finding out whether a students knows a predetermined theory) or divergent (seeking to discover what the students knows). What is required is an approach to assessment that enables both convergent and divergent assessment to be pursued at appropriate times. (para 6.10)

In relation to student motivation:

- There are actions that teachers and schools can take to ensure that the benefits of summative assessment can be had without negative impact on students' motivation for learning. These include
  - promoting and engaging in professional development that emphasises learning goals and learner-centred teaching approaches to counteract the narrowing of the curriculum
  - emphasising learning rather than performance goals with students and developing students' understanding of these
  - providing feedback in relation to these goals
  - developing policies that ensure that the purpose of all assessment is clear to all involved
  - presenting assessment as a process producing results that have to be regarded as tentative and indicative rather than definitive. (para 6.12)
- Students have become very utilitarian in their view of what it is worthwhile to pursue in complying with assessment requirements. This may not exclude enthusiasm for work, and may represent a commitment to the achievement of some long-term goal. (para 6.14)

In relation to feeding back the outcomes of assessment:

- Even when teachers provide students with valid and reliable judgements about the quality of their work, improvement does not necessarily follow. (para 6.15)
- Feedback should be clearly linked to the learning intention so that the student understands the success criteria and standard. (para 6.16)
- Feedback focuses on the task and not the learner, gives cues about next steps, challenges, requires action and is achievable. (para 6.16)
- Improving the quality and extent of feedback may not be incompatible with the use of internal assessment for summative purposes and may become a feature of more sophisticated management of external assessment through the use of ICT. (para 6.17)
- Feedback in relation to summative assessment that does not include comments and information related to the specifics of the work that students have done is of limited value. Students need to be able to take action that will bring particular pieces of work up to the required standard. For students, feedback that is insufficiently focused is of little use. (para 6.18)
- There are links between quality of feedback and students' capacities to manage their own learning. (para 6.19)

The validity and reliability of internal assessment

In relation to what we know about the reliability of teacher assessment:

- The evidence is typically very mixed. (para 7.2)
- Differences between subjects in how teacher assessment compares with standard tasks or examinations results have been found, but there is no consistent pattern. (para 7.32)
- A popular view has been that where qualification outcomes depend on internal assessment they should be accorded a lower status and where high status is required, internal assessment should be limited in its scope and weight and should be tightly specified. (para 7.1)
- Awarding bodies cannot be expected to neglect reliability. (para 7.3)
- Much of the research evidence does not include a consideration of the differences in reliability between assessments based on a single piece of work or many pieces of work, made analytically or holistically, made on the basis of compensation or mastery, and so on. These distinctions are of crucial importance when deciding how an internally assessed component may be specified and which approach is likely to lead to an acceptably reliable outcome. (para 7.4)
- Any set of marks can conceal the operation of all sorts of covert or unintended rewarding, including that which is gender-related, substituting perceived ability as a proxy for achievement, rewarding diligence or tidiness or reflecting expectations of performance. (paras 7.5 – 7.9)
- There is considerable error and bias in teacher assessment of different groups of primary students but the interpretation of correlations of teacher assessment and standard task results should take into account the variability in the administration of the standard tasks. (para 7.32)
- The reliability and construct validity of portfolio assessment where tasks are not closely specified is low. (para 7.32)
- It is wrong to assume that meaning is inherent in the words used to communicate assessment tasks, and that problems can be given ready-made to students. Assessment outcomes must be interpreted in relation to individual contexts and circumstances. Consequently there is a practical limit to the reliability of assessments made through the unmediated delivery of external tasks. (para 7.10)
- A finer specification of criteria, describing progressive levels of competency, is capable of supporting reliable teacher assessment whilst allowing evidence to be used from the full range of classroom work. (para 7.32)
- In general, vocational qualifications studies have suggested that reliability is low and that inconsistent judgements are commonplace. (para 7.19)
- It is not clear whether the disagreements between assessors in studies of vocational or occupational qualifications are larger or smaller than in some essay marking or in some comparable coursework assessment. (para 7.20)

In relation to the methods that teachers use to arrive at assessment:

- Teacher judgements are complex and their interpretation of evidence goes beyond simply matching performance to description. The process requires social construction of standards and the interpretation of evidence involves implicit models of assessment practice. (para 7.13)
- It may be the case that a teacher: starts with the most holistic judgements before referring to the criteria, then shuttles to and fro between the holistic judgement and the criteria in a process of refining the decision; may refer to other students' work in a further process of refinement; may make a trade-off among criteria, because of inconsistencies in performance; may distil standards from common characteristics across a range of students; resolves conflicts that arise when criteria designed to apply to all syllabus implementations are applied in the context of a specific implementation. (para 7.12)
- The equivalent application of standards depends on shared experience and understanding amongst those involved. (para 7.13)
- Teachers who have participated in developing criteria are able to use them reliably in rating students' work. (para 7.32)
- Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others can give. (para 7.32)
- The more teacher assessment is taken out of the control of external examinations, the more they will use their own criteria for making judgements, and the greater the differences will be between these and examination results. (para 7.15)
- In vocational and occupational qualifications assessors may agree on whether candidates are competent but disagree over specific judgements. There are dangers that assessors are easily swayed by the quantity and presentation of evidence rather than by its substance. (para 7.17)
- The variability of the contexts in which competence is tested and displayed means that assessors have to take account of context when judging whether an observed piece of evidence fits a defined criterion, operating with a compensating procedure which itself requires an internalised holistic model and not a simple set of atomised domain descriptors. (para 7.18)

In relation to issues linking reliability with validity:

- External written examinations should be seen as a sub-set of what may be assessed in school-based assessment. (para 7.22)
- Where the assessment is conducted in the same context as the performance, validity is high. Assessment is highly context-specific and one generalizes at one's peril. Assessment tasks and situations lead to best performance when they are concrete and within the experience of the student, clearly presented, seen to be relevant to the learner's current concerns and use situations which are not unduly threatening and where there is a good relationship with the assessor. (para 7.24)
- Teachers will not deem a pupil to be at level 3 without being sure that they have achieved level 2, adapting their interpretations of the domains in order to ensure this. (para 7.26)

- Teachers often sought to build up a picture of the student and his or her understanding, as an essential step in making an assessment judgement. However, this may be a judgement that is of the student's perceived ability, which may result in the teacher overlooking omissions or inferring performances, in order to arrive at what is perceived to be a fair judgement. (para 7.28)
- The growth of authentic assessment has led to teachers and students developing their own tasks and the criteria for assessing them and discussing the interpretation of assessment. Authentic assessment requires that we reconceptualise concepts of validity. (para 7.29)
- Workplace assessment is only valid when it registers accurately the presence of skills which convert directly into occupational competence. Studies of the reliability of assessment are of little consequence if validity is so limited that the competencies acquired are not transferable into the relevant workplace setting. (para 7.21)

Quality assurance and control of internal assessment

In relation to the general requirement for quality assurance and control:

- Control over internal assessment can be provided before it starts, as part of its process, or by looking at its outcomes; just about every combination of these approaches has been used. (para 8.1)
- Teachers and students interact very closely with whatever methods of quality assurance and control are implemented. (para 8.3)
- There may be considerable value in an approach to the quality assurance of assessment that regards this as one aspect of the management of all aspects of quality across an institution. (para 8.5)

In relation to support for teachers:

- Support strategies may provide teachers with materials designed to support their assessment; may provide training in the operation of particular assessments; or may seek to develop teachers' understanding of the principles of assessment. (para 8.7)
- It is possible that too much of the training of teachers has been in the provision of short programmes focused on specific qualifications that do not provide the continuity or depth that is required for real professional development. Yet, high quality provision would be costly and time-consuming. (para 8.10)

In relation to moderation processes:

- Moderation may be seen as an alerting process through spot-checks or statistical methods or occasional scrutinies when danger signs appear. It may not need to be applied to centres that have met a number of quality criteria. (para 8.6)
- Moderation should be part of a process that gives teachers training, includes feedback to teachers and includes moderator training and standardisation. Moderation should be done without knowledge of teachers' marks or grades and sample sizes for moderation should be adequate to ensure comprehensive judgements. Statistical moderation against written papers as criterion should not be used as a sole method, but should trigger inspection. (para 8.11)

- Statistical moderation is poorly understood; teachers see it as inappropriate to use written papers as a moderating instrument for work of a practical nature. (para 8.13)
- Schools often receive very little feedback on the reasons for changes to marks as a result of moderation, and there is therefore little basis upon which teachers could modify their assessment practice. Decisions made within moderation were seen to take little account of the conditions under which teachers worked and how these affected assessments which they make. (para 8.13)
- Any form of statistical moderation using a reference component consisting of a written paper is open to the criticism that the procedure is statistically illogical. Whilst simple statistical moderation is seen to be cheap and objective it is weak in situations where objectives covered in the school assessment are not covered in the test and where entries are small. (para 8.16 and 8.19)
- Statistical moderation generally involves the scaling of marks whilst preserving the rank order of candidates within a group. It is open to manipulation to maximise results. (para 8.17 and 8.19)
- In moderation by inspection there is little direct evidence on the behaviour of moderators or of the processes that they use; and the design of moderation procedures has generally not drawn on detailed fieldwork with teachers and schools. (para 8.22)
- Problems arise when too wide a range of expectations are attached to a quality control process which is insufficiently resourced. These include moderators being required to focus too much on ensuring reliable assessment and not enough on validity, and being unable to provide the required quality control as well as support for the teachers. Moderators differ in the emphasis which they put on these roles. (para 8.23)
- Consensus moderation has been regarded as a component of a golden era of internal assessment, mainly on the grounds that it promoted professional development as well as achieving the required levels of standardisation. (para 8.25)
- In fact consensus moderation was often done on a goodwill basis with many costs actually being borne by schools. It has been used as a method of operationalising teacher control. (para 8.26)
- A distinction should be drawn between procedures which focus on professional development in assessment (agreement trials) and those which focus on making or ratifying decisions about assessment for a specific purpose (moderation). (para 8.27)
- Forms of moderation which are based on quality assurance and result in teacher development and enhanced understanding of the subject matter are to be preferred, but are very costly. (para 8.32)

In relation to processes of verification:

- Verification has often failed to cope with the extensive criticism that the actual outcomes of assessments are not comparable from candidate-to-candidate or centre-to-centre. (para 8.37)

- The external verifier role has steadily increased in complexity, tending to include moderation and guidance as well as verification, and may be close to being almost impossible to manage within the resources available. (para 8.38)
- There is a set of baseline conditions that form an irreducible minimum framework within which internal verifiers impact beneficially upon the quality concerns of their organisations, and also feel adequately supported and professionally developed. These may be summarised: as a coherent accreditation structure; the incorporation of assessment and verifications into the organisation's strategic planning for learning and upskilling; awarding bodies stipulating internal verifiers' tasks; specifying the amount of time that they are expected to spend on verification duties. (para 8.39)

Internal assessment and lifelong learning

Linking assessment and the development of the lifelong learner:

- Little is achieved if teachers are assessing exactly the same things as an external test or examination other than a hiding of the costs and workloads attributable to the examination. This would not be an adequate justification for the introduction of internal assessment. (para 9.1)
- Some commentators allege that some internal assessment for summative purposes has become very stereotyped. (para 9.2)
- Some students have evolved safe identities as learners and adopted strategies that ensure almost minimal compliance with assessment requirements in order to get the qualification that they require. (para 9.6)
- Both formative and summative purposes have an essential role in exerting ever stronger influences as learners progress through education developing learning dispositions that relate to lifelong learning and the development of learning careers. (para 9.7)
- Assessment for learning should develop learners' capacity for self-assessment so that they can become reflective and self-managing. (para 9.8)
- Should students never be required to take decisions related to a piece of work, evaluate findings, reflect on processes and their own performance, set targets or interact with others they may not be equipped to develop learning careers to an extent that will cause them to function as lifelong learners. (para 9.9)

Internal assessment in the national curriculum

Additional evidence from experience with national curriculum teacher assessment includes:

- Teacher assessment continues to operate in the shadow of testing because policy makers have not seen it as an important component of the system and their agenda did not link it to formative and diagnostic aspects of assessment. (para 10.2)
- In the early days of national curriculum assessment teachers undertook the enormously complex assessment tasks largely without central guidance and support and with a poorly articulated purpose for what they were doing. (para 10.3)
- A range of exemplar and supporting materials was provided, together with extended proposals for ways in which departments or other groups of teachers might develop

internal quality assurance. These arrangements have not, for the most part, survived although the materials still exist and are valued. (para 10.5)

- National curriculum assessment is running ahead of theoretical frameworks available to support it. Teachers, faced with competing demands, conflate the formative and summative functions of assessment. (para 10.6)
- The combination of exemplification materials and group moderation is a desirable process for arriving at common standards. (para 10.8)

Some particular internal assessment strategies

In relation to portfolios and performance assessment:

- Portfolios contain evidence assembled by students in order to demonstrate competence and are designed to place students in decision-making roles that involve them in understanding and interpreting the specification for the qualification, and deciding what evidence it will be appropriate to incorporate in the portfolio. In practice assessors normally provide a considerable amount of guidance and sometimes exert almost total control over the evidence-assembling process. (para 11.2)
- The processes by which students manage the compilation of the portfolio are themselves key aspects of learning and the management of learning. The creation of a portfolio potentially entails the management of a work programme, the creation, evaluation, choice and presentation of items to be included in the portfolio and processes of review and decision-making that relate very closely to some aspects of formative assessment. (para 11.3)
- There is a need to shift the emphasis from the collection of evidence in a portfolio to a focus on the analysis and integration of learning. Students should share in the framing of the criteria by which assignments and other portfolio components are assessed. (para 11.5)
- In some situations the marking of portfolios is not by the students' own teachers, but by teachers employed and trained for the process, making them an instrument of external assessment in a very direct way. (para 11.6)
- Some portfolio assessments are not of a sufficient quality to enable comparisons to be made between students for selection or between schools for accountability, though the effects on improved instruction may be considerable. (para 11.7)
- Portfolio assessment places substantial new demands on teachers and schools and there is a need for additional training for preparation and marking. The intention that portfolios should broaden the curriculum and provide a more authentic approach to assessment may be limited by the use of explicit scoring criteria that teachers may interpret as the curriculum goals. In other words, something as flexible as portfolio assessment may be very limiting if not operated in the context of clearly articulated curriculum goals, which operate as the principal point of reference. (para 11.8)
- Whilst it had been said that teachers' assessments of portfolios could only be made acceptably valid and reliable if the tasks and the criteria by which they were assessed were externally supplied and closely controlled, this does create a tension with the need for learners to take some control over the portfolio and for teachers to have some ownership of the assessment process. (para 11.10)

In relation to assessing practical, oral and other performance work:

- The one-off practical examination is said to be a very weak assessment instrument and some of what is assessed in a practical examination may be very close to the cognitive areas covered by written papers. A practical examination is a less adequate measure than teachers' assessments over a number of experiments. (para 11.11)
- The assessment of practical work often involves teachers in observing students working in groups and there are difficulties involved in this when there are many students to be watched and many constructs on which to report. The problem of student-student interaction is little researched in relation to assessment judgements. (para 11.13)

In relation to modular approaches:

- It is possible that internally assessed synoptic assessment will be undertaken under tight conditions of external control, and that it will include the assessment of a wider range of objectives or higher skills than are to be found in many internal assessments. There may be opportunities here for extended pieces of work over a period of time. (para 11.17)

1 Introduction

1.1 This introduction serves only to review the terminology of teacher assessment and to describe the structure of the discussion that follows.

Terminology

1.2 A general description of teacher assessment, leading to some form of classification of its various features, needs to be prefaced by some discussion of the terminology that I am going to use. The very term *teacher assessment* tends to locate the discussion in schools operating a subject-based academic curriculum, and sounds rather alien in other environments. The term *coursework* is normally used where there is a relationship with a terminal external examination, but this can take different forms depending on how the tasks are specified and where in the curriculum they are located. The term *internal assessment* suggests something that is rather more freestanding and more in the control of the teacher or tutor. *School-based assessment*, a popular term in some countries, is narrowly located in schools but is used about assessment that can form the whole of the assessment for a qualification and that is often based on curriculum-embedded tasks under the teacher's control. It was once common to refer to *continuous assessment* that suggested a contrast with a one-off examination (or perhaps a complement to it), with the term most often used in a school context. *School-based examining* is a term that entered the vocabulary with Mode 3 CSE in the UK in the 1960s.

1.3 Parallel terminologies also tend to be located within particular types of assessment, qualification or institution. The differences between *moderation* and *verification* may be more or less understood but the processes are still very strongly associated with particular approaches to assessment. *Examination* and *test* are used interchangeably in some literature, though the latter may be a more generalised term than the former, and may describe the instrument used rather than the process of which it forms a part. The descriptions of an individual as a *teacher* or *tutor* and/or as an *assessor* are strongly embedded in particular settings. Across all of this diversity there are many terms to describe the different purposes for assessment by teachers: *formative, summative, diagnostic, screening, for learning,* and so on.

1.4 We do need to remember that little of the terminology of internal assessment is universal, so that terms such as *teachers' classroom assessment, performance assessment* and *authentic assessment*, widely used in the USA, tend to refer to assessment that is conducted for monitoring, certification and accountability whereas classroom and authentic assessment in the UK are associated with formative purposes. In the UK the term *assessment for learning* has recently been adopted in place of *formative assessment* so as to ensure an unambiguous commitment to learning and motivation, and avoiding formative assessment that is

> "little more than conscientious summative feedback, interim testing and prescriptive target setting designed more for quality assurance purposes than for learning." (Ecclestone & Pryor, 2003: 472).

1.5 There may be other examples of the diversity of our use of language to describe assessments made by teachers. It's not the job of this review to impose a single terminology and, as will be seen later, it is difficult to arrive at a simple universal taxonomy that can operate in all learning environments and institutions, or stand up to

rigorous application to all qualifications. However, the Working Group on 14-19 Reform has been careful to use the term *internal assessment* and to speak of *teachers and trainers* and *learners*, and these do appear to be the most general terms that are available to us. Though I have tried to make my use of terminology as unambiguous as possible I have strayed from the working group's terminology in many places.

1.6 We will also need to be careful to remember the diversity and variation contained within some of these terms. Thus, the term *internal assessment* has a wider application than discussed in this review: it covers formative processes, assessments managed within an institution but used summatively, and assessments conducted within an institution, but controlled externally (such as coursework within an examination). *Teachers and trainers* include those who would call themselves assessors but these are the same people who work in other environments as moderators and verifiers, or whose primary role is that of a supervisor, and whose actions impact very directly on the conduct and outcomes of the assessment. *Learners* are, of course, children, pupils, students, trainees and candidates, with the term used depending on the context in which they are learning.

Structure of this report

1.7 This rest of this review now falls into four main parts, some of which (especially Sections 5 – 11) draw quite heavily on an earlier review (Wilmut, 1999).

Section 2 discusses the development of internal assessment in the summative context, attempting to identify its scope and the main strands of its development and use.

Sections 3 and 4 present a taxonomy of internal assessment and some examples of its use. These sections are intended to provide an illustration of the scope described in Section 2.

Sections 5 – 11 attempt to examine evidence about particular aspects of internal assessments for summative purposes.

Section 12 is a personal view of the directions in which change could now take place.

2 The development of internal assessment

The qualifications context

2.1 In a wide-ranging and penetrating sociological analysis of the development of assessment Broadfoot (1996) locates the roots of our present use of qualifications on a mass scale to the nineteenth-century concern with competence – the need to regulate entry to professions and occupations. Certification conferred status and social advantage, and the assessment methods themselves acquired a similar high status with an increasing role in selection as well as certification. Broadfoot and others have argued strongly that assessments have thus become agents of social and individual regulation and control (Broadfoot, 1984).

2.2 In order to cope with the numbers of candidates, and because of the relative ease with which quality could be controlled, written tests became progressively more common, gradually displacing traditional systems of apprenticeship (in which assessment of competence was done on the job, often in relation to local or company requirements); these became increasingly restricted only to craft occupations before almost disappearing in the late twentieth century. More importantly, the use of such formal methods of assessment on a mass scale resulted in the emergence of the syllabus and the concept of the curriculum that later formed the basis for the conduct of mass education and training.

2.3 The association of the test or examination with access to high status professions and occupations was reinforced in the early twentieth century by the emergence of concepts of innate ability and of psychometric testing that itself came to form a critical part of the process of selection for all 11-year-olds. Broadfoot emphasises the attractiveness and power of this use of tests, and that

> "... the scientific, 'objective' nature of such tests, their proven predictive power and their measurement of a characteristic believed to be as inborn as eye colour, meant it was almost impossible for the recipient to reject the diagnosis." (Broadfoot, 1996:35)

2.4 Examinations and tests have come to be regarded as objective, fair, equal for all candidates, and having high reliability. They have undoubtedly been managed and delivered very successfully over a long period of time, and have achieved considerable status in their own right, although the details of examination processes have traditionally been hidden. Moreover, their status has been strengthened by the early belief that they provided motivation for learners to work and subsequently with the belief that they ensured that educational standards were maintained. This has led to the point where these issues have often been a primary focus for the public debate of education and training issues. Further reinforcement of their status has come from the more recent use of examination results as indicators of educational system performance, and through the deliberate strategy of manipulating curriculum change through the management of assessment processes – known here as assessment-led curriculum development and in the USA as measurement-driven instruction (Torrance, 1995).

2.5 Thus, by 1950 there was a wide range of tests and examinations that were almost wholly external in character, operated by examining boards, professional bodies and universities. Few of them were accompanied by any form of internal assessment except

in a range of industry-based apprenticeships, where on-the-job competence continued to be highly valued and was assessed directly by practitioners, frequently working within occupation-specific or industry-specific standards that had high status within their own sectors.

2.6 We will look first at the broad trends in the development of internal assessment in qualifications in the period since 1950, and then review some specific issues that are likely to be central to the present discussion. At this point it is important to stress that there has been no *planned* development of internal assessment over this period. It has been a component of a number of initiatives and a by-product of a range of provisions but has rarely, if ever, been viewed as a whole. Until the last few years there has been almost no perspective on internal assessment that has embraced all general, vocational and occupational qualifications and there is still only a limited understanding of the relationships between internal assessment for summative purposes within qualifications, internal assessment for summative purposes within the national curriculum and assessment for learning. There are risks in attempting to treat internal assessment as an integrated phenomenon and in generalising from one context to another.

Growth of internal assessment in school examinations

2.7 In the 1950s there were signs of impending change to school examinations, much of which would accompany the move away from selection for secondary school at 11+. In a debate that foreshadowed the discussion of authentic assessment it was being argued that the validity of external school examinations would be enhanced if some of the objectives were to be assessed by class teachers; the alternative was

> "... leaving unassessed the ephemeral *processes* if investigating, analysing and problem-solving which proponents of authentic assessment would claim to be of equal if not greater importance that its products." (Torrance, 1995: 46-7)

2.8 Moreover, it was said, such a move would bring summative assessment closer to the formative assessment of the classroom, enhancing the latter, with examinations that better reflected the full spectrum of learning in each subject and, perhaps, made more explicit what was to be assessed. The negative aspects of the backwash on the curriculum would be reduced and teacher expertise in assessment enhanced, arguments later discussed by Kellaghan & Greaney (1992), Murphy & Torrance (1988) and Akyeampong & Murphy (1997). This debate extended considerably in the 1960s with the newly introduced CSE, where it was proposed that teacher control over examinations would extend in various ways: there would be internally assessed components in examinations in Modes 1 and 2, and all Mode 3 examinations would be internal[1]. The origins of these proposals went back to the Norwood and Beloe reports, but now the Schools Council Examination Bulletin 5 laid the foundation for the Certificate of Secondary Education (CSE) structure, and described the basis upon which the examinations could be moderated. Many of the procedures it described were new to schools (such as agreement trialling), but the innovations had a profound effect on British examinations of all types (Schools Council, 1965).

[1] Mode 1 examinations used board syllabuses, with examinations set and conducted by the board. Mode 2 examinations used approved school-devised syllabuses, but with examinations set and conducted by the board. Mode 3 examinations used approved school-devised syllabuses, and school-devised and conducted examinations, subject to appropriate quality control procedures.

2.9 The delivery of CSE across the country did not follow a single pattern, with the regional boards each adopting their own approaches, in agreement with their Local Education Authorities (LEAs) and schools. The take-up of Mode 2 and Mode 3 systems never exceeded a quarter of all subject entries (Mode 2 was never exploited to any great extent) despite the considerable boost to CSE that followed the raising of the school-leaving age in 1974. The examinations were often of a quite traditional type, though considerable quantities of internal assessment appeared, described either as *coursework* or as *continuous assessment* (Rogers, 1974).

2.10 Mode 3 attracted strong support from teachers who wanted to innovate on the basis of a clear view of their curriculum goals, and it survived as an approach until about 1990, though increasingly threatened by the imposition of general and specific criteria for the new General Certificate of Secondary Education (GCSE) examination. Straughan & Wrigley (1980) commented that

> "… in Mode 3 the link between values and evaluation is seen very clearly – if the teacher makes his own assessments on his own terms and in his own school, then his values and his evaluation are closely linked. The intensity of effort needed to achieve this integration is one reason why Mode 3 has always been the preoccupation of a minority. It is easier to work to a Mode 1 external syllabus…"

and also said of external moderation that it "… takes account of the values of the individual teacher or school"

2.11 Cohen & Deale (1977) provided a comprehensive review of teacher assessment in examinations in the 1970s. They identified three approaches to teacher assessment in CSE, namely

- where teachers assessed the same skills and abilities as were assessed by the external paper(s) and both were used to generate a result that was said to have enhanced validity and reliability of the qualification as a whole
- where teachers assessed aspects which the external papers could not, thus justified as enhancing validity
- where teachers assessed internally but also marked external papers, supported by moderating/standardising groups.

They contrasted CSE practice (where most examinations had some teacher assessment) with General Certificate of Education (GCE) practice (where there was still very little). The weighting for teacher assessment in CSE Mode 1 examinations varied between boards but (across main subjects) was generally at between 30% and 60%.

2.12 The stance of the GCE boards varied but was generally cautious: some had begun to explore the introduction of internal components before CSE began, and continued to develop these approaches, though quite slowly. Others continued to argue the merits of external examination systems, which needed to be "... scrupulously honest and impartial and seen to be so", while taking a sideswipe at Mode 3 which is "... only loosely controlled by an examining board". In addition "... marking has to be as reliable as possible" and "... standards throughout the country [must be] comparable" (University of Cambridge Local Examinations Syndicate, 1976). Many GCE board publications of the time regarded external written examinations as the 'normal' method, and internal assessment was seen as a rather risky venture, although the boards did meet together to

discuss the issues in a major seminar in the early 1980s (Associated Examining Board, 1981).

2.13 There were proposals for extending CSE (Keohane, 1979) by developing a qualification to follow it, and having a similar format, with a recommendation that this should link more explicitly across from academic subjects into the vocational curriculum and to occupational qualifications. In some ways this anticipated the growth of General national Vocational Qualification (GNVQ), but it never got far off the ground though it did fit quite closely with the way that some BTEC programmes operated, with a large general educational component pursued in a vocational context and linked to a range of generic skills.

2.14 The more significant change came with the Waddell report that sought to draw together the experience with GCE O level and CSE, and from a range of feasibility studies conducted in the early 1970s, to develop a common system of examining at 16+ (Waddell, 1978). In the report Waddell suggested that

> "the introduction of a common system is likely to involve more teachers in responsibility for assessment of their pupils' performance, and wide reliance will need to be placed... on course assessment and practical tests…"

Much later Broadfoot (1996) was to comment that Waddell's assertions that a common examination system was feasible and that this should be run by 4 regional consortia was a major factor in legitimating a much greater degree of control by central government than had previously been the case, and that this led ultimately to the use of assessment and the examination system as a mechanism for achieving greater control over the curriculum.

2.15 The early 1980s saw a reduction on the emphasis on internal assessment, in the run-up to the introduction of GCSE. Bowe & Whitty (1984) took a particularly gloomy view of the prospects, and Nuttall (1984) correctly predicted that the draft national criteria for GCSE – to which all examinations were eventually to be required to adhere – would kill off Mode 3 because of a wish, on the part of the government, to establish central control and regulation of all school examinations. Hargreaves (1982) was also pessimistic, suggesting that the practical success of school-centred innovation "is balanced on a knife-edge of uncertainty", and observing that the majority of teachers were excluded from many important areas of decision-making.

2.16 Although the Secondary Examinations Council (SEC) – set up to oversee the introduction of GCSE – initially adopted a rather traditional GCE way of speaking about the new examination (SEC, 1985), it had become rather more open-minded three years later (SEC, 1988), when a good deal of advice on internal assessment was offered to schools, drawing on the work of schools, colleges and LEAs. Coursework became a significant component of GCSE, only to be cut back arbitrarily in 1991, in the muddle that surrounded the attempts to reconcile GCSE with the requirements of the national curriculum and its assessment (Daugherty, 1995).

2.17 GCSE was regulated through the use of general and subject-specific criteria, to which all proposed syllabuses had to conform. The general criteria (appearing in their original form in 1986) introduced into the discussion of teacher assessment the idea of 'fitness for purpose' and said that "any good scheme of assessment and moderation must … not be allowed to dominate the educational aims or inhibit good teaching and learning practice". Wood (1991) took the view that the commitment to the use of teacher

assessment where the validity of the assessment demanded it meant that it could not be bargained away, however awkward it was to manage and suggested that

> "… in effect, school-based assessment becomes an engine for enriching the curriculum as it is delivered".

2.18 Developments at A level had, meanwhile, proceeded rather more sedately. There was generally less internal assessment at this level than in 16+ examinations, although several of the examinations boards collaborated on a range of innovations in the 1960s, 70s and 80s that explored a variety of approaches to project and practical work (such as the Nuffield sciences, projects in History and Geography developed by the Schools Council, and the Wessex Project[2]). As with GCSE, A level became more closely regulated through the 1990s, with codes of practice applied to various key processes in its operation, and mergers between examining boards, together with government pressure, reduced the number of courses available.

2.19 At the same time there were several experiments with modular approaches and a significant discussion of the mechanisms for creating and using a modular and/or unit-based curriculum. Much of this was linked to hopes (fuelled by work on TVEI – Technical and Vocational Educational Initiative – projects around the country) that it would be possible to 'bridge the academic and vocational divide', allowing students the opportunity to build programmes that suited their needs. This was sometimes characterised as a free-for-all but most of the schemes that were tried out used quite restrictive rules by which modules could be linked, and the most conservative approaches were no more than the chunking of a full A level programme with, perhaps, the addition of some optional modules. It was common to find internally assessed components in many modules.

2.20 Almost all post-16 programmes are now modular in structure, although the more radical structures that allow exchanges of modules between programmes have found little support. The present split into AS and A2 and the various provisions available within GNVQ and Advanced Vocational Certificate of Education (AVCE) depend entirely on the existence of modules, and systems for accumulating or combining unit credit. These schemes do offer the possibility of retaining credit for a limited period, although almost all programmes are followed continuously. Some reorganisation of educational provision has resulted in a much larger proportion of A level programmes being offered in Further Education (FE) or sixth form colleges.

Developments in assessing vocational and occupational qualifications

2.21 The diversity and number of vocational and occupational (sometimes known as technical) qualifications has already been noted and these were gradually becoming more and more widely available, particularly as further education colleges began to expand their range of provision (that gradually included a larger and larger provision of GCSE and, particularly, A level programmes). Wolf (1995) discusses the diversity that existed until the early 1980s, the large number of awarding bodies involved (although

[2] This was one of a number of projects that grew out of the TVEI initiative, launched in 1982 by the Manpower Services Commission and operated through projects run in partnership with LEAs. TVEI did not finally end until the mid-1990s though, by that time, many of the innovative approaches to the use of internal assessment had either foundered because of the application of the general and subject-specific criteria of the GCSE, or had become absorbed into more widely available provisions from the examining boards.

three large ones predominated) and the near-collapse of the apprenticeship system, so that

> "What existed up until the 1980s was thus a nascent tripartite system but one in which only academic A levels were a coherent part of the education system, and in which other vocational and technical awards suffered from the complexity of the awarding system and the almost total lack of understanding, by higher education and employers, of what the different awards meant and how they might relate to each other."

This led to reform that culminated, in the mid-1980s in the emergence of National Vocational Qualifications (NVQs), using a competence-based approach, with reformers seeing these as more enabling, less elitist and based directly on the requirements of industry. The development of the NVQ structure led, in turn, to the development of GNVQ, initially using a similar outcomes-led model, and designed to

> "… offer a meaningful qualification to non-traditional learners with relatively poor prospects for education and work" in which "… summative evidence of achievement and formative feedback on progress would feed organically into each other and that students would be active agents in negotiating these processes." (Ecclestone & Pryor, 2003).

Subsequently, pressure over issues of standards and technical difficulties with the GNVQ model

> "… created new imperatives to regulate and standardise teachers' assessment decisions through prescriptive activities regulated by the Qualifications and Curriculum Authority. These came to exert a powerful influence over teachers… shaping their formative feedback and summative expectations in particular ways". Then "… the assessment regime encouraged a subtle self-regulating acceptance of its purposes, practices and effects by teachers and students alike."

2.22 Regulation and standardisation came through a number of routes and changed over the 1990s. Training units for assessors, internal and external verifiers and others, initially developed for NVQs, were adapted and increasingly required for those working with GNVQs. On the basis of recommendations made by Capey (1995), Dearing (1995) and others an increasing level of external assessment was inserted and the internal assessment processes were simplified in order to reduce teacher workloads. Mechanisms for internal and external verification and for moderation (discussed in more detail later) were made more explicit and there was a considerable investment in providing support materials that would extend teacher expertise in setting assignments and conducting assessments. Changes to Advanced GNVQ to make it operate rather more like A level was presented as a means of better integrating it and making it more credible, but left many practitioners feeling that its appeal had been diluted for the group of learners for whom it had been intended.

2.23 In the background, discussions about the validity and reliability of assessments in NVQs and GNVQs became sharper in the 1990s, focused on whether the competence model was appropriate, whether the conduct of assessment was valid, and whether the outcomes were reliable. Wolf (1999) has provided a critical discussion of the assessment of vocational qualifications in the UK over the last decade, and the relationship with teaching and learning, also including a discussion of parallel issues as they have occurred in some other countries. Having traced the development of NVQs

and GNVQs, and the relationships between these and other vocational and academic qualifications, she explores the notion of competence and the collection, by the individual, of evidence which demonstrates aspects of competence in a particular vocational or occupational area; these are related to highly detailed (some would say fragmentary) performance statements that, taken together, have led to a very complex, cumbersome and bureaucratic assessment system, dominated by paperwork that records, explains and justifies a large number of separate assessment decisions.

2.24 Three important points, relating to internal assessment, emerge.

- There has often been a belief that making the statements which specify competence more specific and more exact will result in greater clarity and more valid and reliable assessment; this has not been the case. Re-writing has resulted in more complexity and less clarity, as well as further obscuring the overall concept of competence in the occupational or vocational area.
- Assessors tend to develop and use concepts of overall competence; this does incorporate some trade-offs between aspects of competence, and there is limited evidence of whether these are comparable from assessor to assessor. Other work suggests that views of 'custom and practice', drawn from an occupational sector or workplace, are sometimes substituted for an exact application of the standards (Wilmut, 1994a), and Ecclestone & Hall (1999) have suggested that concepts of a 'standard' derive from a complex and ad hoc mix of experience, liaison with colleagues and others, and some informal sharing of assessment judgements. They tend to apply these notions regardless of a particular quality assurance system.
- Whilst teachers and assessors have often expressed strong support for the competence model used in NVQ and GNVQ, and most seem to have developed their own strategies for managing the assessment (Wolf, 1998), the Capey review of GNVQs in 1995 and Further Education Development Agency (FEDA) studies from 1994-7 all comment on high assessment workloads, said to be unacceptable by a proportion of teachers. Wolf suggests that standardising these processes is very difficult, though simply standardising the outcomes may also be almost impossible.

2.25 There were also Office for Standards in Education (OFSTED) and Further Education Funding Council (FEFC) criticisms of some aspects of learning and assessment in GNVQ (cited by Wolf), suggesting that students had become "... hunters and gatherers of information" in order to satisfy the evidence requirements. Two more recent studies support and extend this view. Ecclestone & Pryor (2003) describe how GNVQ practices in colleges interacted with students' dispositions to learning that

> "… formed 'horizons for action' that influenced in powerful but subtle ways teachers' and students' engagement with formative assessment. For example, horizons for action turned activities such as oral and written feedback on students' work, reviewing progress and setting targets, and classroom questioning, into strong compliance with the assessment criteria from teachers and students alike. This compliance was a 'pragmatically rational' response to the conflicting pressures they were under. Yet this compliance was not merely cynically adopted… all but one teacher and all students in the study internalised the rationale for it and came, over two years, to promote it as an educational goal in its own right." (p.479).

2.26 They also describe how students were able to articulate the specific language of the assessment specifications (which had been seen as an important aspect of their

ownership of the assessment process) and could use these in a way that legitimised certain knowledge, values and norms whilst giving them the perception that they were autonomous learners. They suggest that

> "… students... evolved safe, positive identities as learners… aiming low, playing safe and working informally with friends was therefore crucial to a new identity"

whilst they also pursued high grades and the progression that they wanted. These findings mirrored those from an earlier study, based on interviews with learners from primary schools to sixth forms and colleges, in which GNVQ and (to a lesser extent) A level students articulated their expectations of learning programmes, assessments set by teachers, feedback and their qualifications and progression goals (Weeden & Winter, 1999).

2.27 During the 1990s there was considerable pressure on NVQs and GNVQs to increase the 'rigour' of the assessment and, in particular, to ensure some assessment (usually through an external component) of underpinning knowledge. This is an often-cited pattern of provision in some other parts of Europe (Wolf cites France, Germany and the Netherlands as examples) where syllabuses have a more traditional pattern, occupational or technical courses are delivered in education or training establishments rather than being designed solely for the workplace, and assessment is by a mix of paper-based and practical examinations. In the course of considering the use of external assessment in NVQs Johnson *et al* (1995) discussed a range of issues, some of which would involve teachers in working with materials which were externally devised or specified, but incorporated into the normal activity of the NVQ. They saw the use of an external assessment as, primarily, enhancing credibility, but noted its potential as a moderating instrument, defining its scope as follows.

- The assessment process should contain at least one component which is common to all candidates.
- The common component should be seen to be assessed independently of local assessors.
- The component must encompass a significant part of the competence being assessed.
- The component must not add to the assessment burden.

The emergence of authentic assessment

2.28 The growth of internal assessment in UK examinations was accompanied by an increasing interest in the assessment of performance, and assessment processes and procedures that more accurately reflected the contexts of desired learning (Torrance, 1995; Madaus & Kellaghan, 1993). This interest came at the same time as the greatly increased use of assessment as a tool of educational policy and control, and the convergence of the various strands of work on assessment for the national curriculum, school examinations, assessment for vocational qualifications and assessment in higher education.

2.29 Authenticity in assessment is not, however, achieved by just injecting a bit of internal assessment into an external examination. In a review of the assessment of students in the United States Darling-Hammond (1994) discusses the introduction of authentic and performance-based assessments, how these may be devised, and how they may relate to the conduct of learning programmes. She also considers the relationship between these

approaches to assessment, the resources for schooling, the expertise of teachers, and the expectation that standards of education will be raised. She concludes:

> "I have argued here that, for all the promise of more authentic and performance-based forms of assessment, their value depends as much on how they are used and what supports for learning accompany them as on the new technologies they employ. Changing assessment forms and formats without changing the ways in which assessments are used will not change the outcomes of education. In order for assessment to support student learning, it must include teachers in all stages of the process and be embedded in curriculum and teaching activities. It must be aimed primarily at supporting more informed and student-centred teaching rather than at sorting students and sanctioning schools. It must be intimately understood by teachers, students, and parents, so that it can help them strive for and achieve the learning goals it embodies. It must allow for different starting points for learning and diverse ways of demonstrating competence. In order for schooling to improve, assessment must also be an integral part of ongoing teacher dialogue and school development.
>
> In short, we must rethink the uses of assessment, since we have entered an era where the goal of schooling is to educate all children well, rather than educating a 'talented tenth' to be prepared for knowledge work."

2.30 This discussion clearly leads us into two areas highlighted later in this review: the relationship between assessment and learning and the reliability and validity of internal assessment. Darling-Hammond's position suggested that we should aim for *reasonable* consistency and *reasonable* reliability, but not necessarily seek to maximise these, and not necessarily try to make internal assessments as reliable as written components. The issue is one of purpose: if the primary goal is to maximise reliability then internal assessment might be an inappropriate tool. If the primary goal is to "… harness a powerful tool for learning" (Gipps, 1994) then internal assessment may be essential.

2.31 There are, of course, some risks that we will highlight later on: internal assessment has often been too dependent on paper-based evidence (which is a characteristic of externally controlled assessment), and it has long been recognised that there is a danger that coursework can easily degenerate into a teacher-controlled mini-examination (Wood, 1991). Moreover, as Black (2004) has suggested, there is a significant difference between teachers' use of external assessments over which they have no control and their use of external assessments that they choose and which they can entirely control.

2.32 While the growth in authentic assessment has been an important force for change, Broadfoot reminds us that other forces have been at work. She rejects the notion that the adoption of internal assessment is because of a greater trust of schools. Rather, it is because of the proliferation of routes through education: whilst there is a single route the assessment process for selection must be

> "invested with as much apparent objectivity, ritual and formality as possible so that the results and the failure which they imply for many candidates are accepted." (Broadfoot, 1996: 46)

Many routes tend to include a broadening of opportunities and the spectrum of what is assessed, and teacher assessment becomes not only essential but acceptable (though within limits).

Summary

2.33 In relation to the development of public examinations:

- Examinations and tests have come to be regarded as objective, fair, equal for all candidates, and as having high reliability. They have been managed and delivered very successfully and have achieved considerable status that has been strengthened by the beliefs that they provide motivation for learners to work and that they ensure that educational standards are maintained. (para 2.4)
- There has been no planned development of internal assessment over the last 50 years – it has been a component of a number of initiatives but has rarely, if ever, been viewed as a whole. Until recently there has been almost no perspective on internal assessment that has embraced all general, vocational and occupational qualifications. (para 2.6)
- There is still only a limited understanding of the relationships between internal assessment for summative purposes within qualifications, internal assessment for summative purposes within the national curriculum and assessment for learning. (para 2.6)
- There are risks in attempting to treat internal assessment as an integrated phenomenon and in generalising from one context to another. (para 2.6)
- Traditionally it has been argued that the validity of external school examinations would be enhanced if some of the objectives were to be assessed by class teachers (para 2.7) and that such a move would bring summative assessment closer to the formative assessment of the classroom, enhancing the latter. (para 2.8). It has also been suggested that internal assessment for summative purposes involves teachers in taking responsibility for assessment of their pupils' performance. (para 2.14)

2.34 In relation to vocational and occupational qualifications:

- Vocational and occupational qualifications have involved the collection of evidence of competence that is related to highly detailed and fragmentary performance statements. This led to a very complex, cumbersome and bureaucratic assessment system, dominated by paperwork. (para 2.23)
- There has often been a belief that making the statements which specify competence more specific and more exact will result in greater clarity and more valid and reliable assessment. This has not been the case: re-writing has resulted in more complexity and less clarity, as well as further obscuring the overall concept of competence. (para 2.24)
- Concepts of a 'standard' derive from a complex and ad hoc mix of experience, liaison with colleagues and others, and some informal sharing of assessment judgements. (para 2.24)
- Assessment activities have turned into strong compliance with the assessment criteria. This was a pragmatically rational response to the conflicting pressures they were under that was promoted as an educational goal in its own right. (para 2.25)

2.35 In relation to authentic assessment:

- Authenticity is not achieved by just injecting a bit of internal assessment into an external examination. (para 2.29)

- If the primary goal is to maximise reliability then internal assessment might be an inappropriate tool. If the primary goal is to harness a powerful tool for learning then internal assessment may be essential. (para 2.30)
- There is a significant difference between teachers' use of external assessments over which they have no control and their use of external assessments that they choose and which they can entirely control. (para 2.31)

3 Developing a taxonomy for internal assessment

The basis for a taxonomy

3.1 Before going on to discuss some specific issues of internal assessment for summative purposes it may be helpful to attempt to create some sort of formal structure which describes its essential characteristics – I will call this a taxonomy although it may not be capable of accurately classifying all of the forms that internal assessment has taken. The difficulty is that internal aspects of summative assessments have often developed in a piecemeal fashion without benefit of over-arching rationale or method. There are some exceptions to this: NVQs and teacher assessment within the National Curriculum, for example, but many of the uses of internal assessment for summative purposes have evolved in response to particular immediate requirements in a subject, building on previous experience rather than implementing a rationale. It is possible that there will be more than one taxonomy that will describe the range of this experience, and likely that all such taxonomies will be approximate and inexact in places.

3.2 In the present context a taxonomy will be restricted to internal assessment used for summative purposes although (as will be discussed in the course of presenting the dimensions of the taxonomy) there are considerable overlaps with assessment for learning. What is important to recognise is that this is not a 'natural' taxonomy nor an attempt to describe all internally managed and conducted assessment, nor a set of pigeonholes within which all examples of internal assessment can be slotted and explained. However, the taxonomy is probably fairly universal (that is, not limited to experience in the UK) and is probably not wholly confined to general qualifications. It may, for example, relate to aspects of assessment of key skills and is intended to include vocational and occupational qualifications. It may also relate to national curriculum assessment where teacher assessment for formative and summative purposes most obviously shade together.

3.3 The taxonomy does reflect the social, educational and political climate of the early 21st century, in which very high stakes attach to many qualifications which are used to recognise achievement (a certification process), for selection and for monitoring system and individual performance. These are conditions that are unlikely to change in the UK elsewhere in the world in the foreseeable future.

Dimensions of this taxonomy

3.4 The taxonomic structure envisaged here is composed of four major *elements*; these are not entirely independent of one another but appear to allow us to describe most forms of internal assessment for summative purposes that have been used. The elements refer to the construction, process and quality assurance of the internal assessment. Within each element there is a series of *sub-elements* that provide us with its characteristic features. It originated in a description developed by Harlen (2004) that specified a continuous 2-dimensional space with axes representing the degree to which the assessment tasks are specified and the degree to which the assessment criteria are specified. She identifies four main approaches within this space: high task/high criteria specificity, low task/high criteria specificity, low task/low criteria specificity and high task/low criteria specificity. Both of Harlen's dimensions are presented as continuous scales and they have been used as two of the elements of this taxonomy, although the breaking down

into sub-elements has created a more complex structure that is not easily represented by continuous scales.

3.5 An early version of this taxonomy added two further elements and Newton (2004, private communication) suggested one more over-arching element and some changes of emphasis that have been incorporated. These additional elements are more naturally categorical in nature than the dimensions of Harlen's original model and most of the elements and sub-elements emerge in the discussions in later sections of this review.

***1** Overall assessment design*	The relationship between the teacher-assessed part and the whole of the summative assessment; this covers three aspects: contribution of the teacher-assessed part to the whole, the domains to be assessed and the nature of the relationship between the results from the parts	
	Contrib-ution	This may be seen as a continuous scale where the teacher-assessed component may form, for example, • the whole of the summative assessment or the whole of the assessment within a unit or module; this may involve the use of one or a number of tasks that are assessed within the component • a large proportion of the summative assessment (say, over 50%), in which case the other component will fulfil a specific role in, for example, assessing aspects of underpinning knowledge, or a domain that is not economically or effectively assessed through coursework • a minor part of the whole assessment, perhaps 20%, where it is restricted to those aspects that are poorly suited to assessment through the written examination
	Domains assessed	The domains assessed relate closely to the size of the teacher-assessed component. The approach reflects the status as well as the relevance of the teacher assessment. • all domains may be assessed • the domains assessed that are best located in work in the classroom or in extended work, or involve process or ephemeral evidence (such as performance) • only those domains that cannot be assessed through the written examination are assessed
	Assessment occasions	Assessment decisions may be made • on a single occasion; as in the observation of the performance of a task • intermittently, assessing a range elements of an extended task or taking several observations of the same element, according to a specific timetable • continuously, normally arriving at an overall judgement (say, about level) based on many observations, normally chosen naturally (such as might be dictated by the context of a learning programme)
	Assessment role	An internal assessment may have different roles within the qualification in which it is operating and results may be reported in a number of ways. • it may stand by itself as where internally and externally assessed components were reported separately – whether it is legitimate or useful to report an assessment component because of the way in which the assessment has been conducted (in contrast to the domain being assessed) is a matter of debate • it may be confirmatory, requiring a demonstrable relationship between the domains assessed internally and externally, as where an external assessment is used to moderate an internal one • it may be additive with internal and external components contributing, according to pre-specified weights, to an overall result – hurdles might be applied making no grade possible without the internal component or with the internal component only able to enhance the overall result

2 Specifying the tasks	Tasks (activities, projects, assignments, etc.) are done by the student and assessments may be based upon them. We may identify three elements that define the nature of each task.	
	Task origin	The task may originate in activities being done in pursuit of a particular aspect of a curriculum or programme of work or as part of a workplace activity. At the other extreme it may be prescribed externally. Some points along this continuum may be, roughly, where • there is no constraint: teachers and/or students choose the topics and their extent or these arise naturally within the programme of work; a high degree of control over the conduct of a task may rest with the student • teachers choose tasks within a framework of criteria that are embedded in a programme or qualification specification • teachers choose embedded tasks but have their choices vetted externally before students can start work on them • teachers and/or students select from an externally provided list that may be either fixed or variable within the framework of the programme of study • teachers and students must use a prescribed topic but may choose the approach that they use • the topic and approach are prescribed (this may also describe external tests or materials supplied in advance for such tests)
	Task duration	The duration of a task relates to its importance or defines its contribution to an overall summative assessment and hence its weight. Tasks might be • long, such as a research project or extended assignment undertaken for inclusion in a portfolio of work • relatively short, but demanding a significant input and carrying a significant weight, such as an extended examination question or a critical workplace task • short, such as a multiple choice or short answer question or a small component of a workplace activity
	Task type	Tasks can involve a wide range of activities, such as • a project or investigation • fieldwork • practical work • the preparation of a report or a presentation • a performance
	Evidence compilation	Evidence for assessment may be collated in a variety of ways that include, for example, • portfolios of work or a compilation of samples of what has been done • the presentation of a single major report or piece of writing • a photographic or video record of a presentation or performance • artefacts and designs • witness statements or self-assessments

***3** Assessment processes*	Here are dimensions along which the assessment processes may be described.	
	Assessment scope	In relation to each domain being assessed • judgments may be made on a mastery basis (pass/fail) • there may be compensation between performances on sub-domains
	Assessment judgement	Assessors may be required to make a variety of types of judgement, such as • selecting a best performance or an average or typical performance • conducting a review across many pieces of work using aggregation rules • making a holistic or impressionistic judgement across a range of tasks • making an end-of-year judgement • using an analytic marking scheme supplied externally • using internal marking schemes or assessment criteria
	Assessment records	These may be compiled as • detailed notes of processes and activities • tick lists of tasks accomplished • marks of specific written outputs • levels or grades achieved on a specific tasks • a mark on an interval scale, for aggregation • a rank
	Assessment support	Internal assessment might require support of several types. • training programmes for assessors • freestanding information or exemplification packages (which may also be used in training programmes) • institutional or local group agreement trialling • periodic feedback • an advisory service to be available for consultation

***4** Quality Assurance*	Internal assessment for summative purposes is unlikely to operate without some form of quality assurance. Mechanisms are of two types that may both have internal and external components and some choice of methods.	
	Quality Assurance approaches	We can distinguish two broad approaches relating to • systems assurance that ensures that there are sufficient mechanisms in place to guarantee that the assessments have been adequately quality assured (or are of the required validity and reliability); this may require the existence of written procedures for the conduct of assessment, staff training provision and internal mechanisms for moderating assessment judgements; this approach may be backed by a risk assessment procedure that allows some centres more freedom than others • the inspection of outputs to see that the assessments that have been made are acceptable; this may be the more direct and demanding quality assurance for a particular assessment but may not directly address the quality of the systems that have generated it
	Quality Assurance components	Quality assurance may be conducted • internally, under control of the institution where the internal assessment is undertaken and/or • externally, on behalf of an awarding body
	Quality Assurance methods	Quality assurance procedures may involve • checking that appropriate systems are in place – a remote procedure • observing those systems in action – an inspection procedure • checking the outputs under the same conditions as when the original assessments were made • checking some or all of the outputs remotely and (usually) at a later stage Each of these methods may be based in a peer quality assurance process (such as in a consortium moderation) or by an agent of an awarding body.

3.7 There may be other ways of describing internal assessment but this taxonomy appears to allow for most alternative descriptions, some of which appear in the brief for this review. A few of the alternatives deserve some particular comment.

Task-based versus observation-based

Superficially this appears to be a distinction between the deliberate setting of tasks for summative assessment purposes as against assessment conducted 'opportunistically' – that is, in the course of everyday learning activities. To some extent this appears to be covered by the *Specifying the tasks* element above in that this allows the possibility that teachers might make selections of the tasks that they will assess. It does have another aspect, however, in that an observation-based approach may suggest the assessment of attributes and/or outcomes that are not included in the assessment domains or criteria required for the summative assessment. This is not excluded in the classification above, though it might be seen to create difficulties in the context of a summative assessment process.

High stakes and low stakes

When we speak of 'high stakes' assessment we are describing the purpose to which the interpretation or application of the assessment outcomes are being put, and not (in principle) the assessment itself. Of course, if the stakes are high this will have a backwash onto the assessment itself, principally on the choices about the form that this should take, but probably also on its conduct. For example, a so-called high stakes assessment may be rehearsed and conducted under tightly controlled conditions, and may induce anxiety in the assessor and the student. While we must takes all these factors into account, it is not obvious how we may use the high/low stakes distinction as a basis for classification – it is, rather, a feature of the environment in which the internal assessment is designed and then the environment in which it is conducted, all of which are reflected by the position that it occupies in the taxonomy.

Teacher control and learner control

If the present review were considering the whole conduct of assessment by teachers, for all purposes, this classification would have greater potential to be incorporated as an element in its own right. The situation with internal assessment for summative purposes appears to place teachers and students 'on the same side' with the external agencies on the other. The extent to which either of them exerts useful or significant control over the choice and conduct of a task or assignment depends on decisions made externally, and expressed through the way in which the tasks and assessment processes have been specified (elements 2 and 3). Whilst there is the possibility that actions by a teacher may limit the degree of control exercised by the student, it does seem more generally useful to use the classification above, in which the relationship between internal and external control is the principal underlying dimension of interest.

The nature of learner decision-making and control might need to be expanded within a taxonomy such as this: it would only be in the most flexible and open-ended task prescriptions that a student would be able to reflect and review, leading to decisions about the next steps to be done within the task (such as in a design problem, or in some occupational contexts), and I have discussed something of the relationship between this and lifelong learning in a later section.

Degree of formality in the assessment

This potential dimension seems to be consequential rather than controlling. Internal assessment for summative purposes is likely to be relatively formal, but this derives from, for example, the way in which the task is specified and the requirements of quality control are to be met. The assessment procedures will also have some formality (in that they will be conducted deliberately and according to externally imposed rules and criteria).

Degree of assessment expertise

The extent to which teachers and assessors are able to conduct assessment is certainly a consideration in deciding whether to introduce internal assessment, and what forms it should take. It has been incorporated here as a sub-element of assessment purposes and processes and as an aspect of quality assurance. Choices about assessment methods will need to take account of whether assessors have the resources and competence to use them.

Recording

This appears to relate to the nature of the criteria used for making assessment judgements, and has been included.

Accountability

This may have been covered in the same way as the high stakes/low stakes distinction discussed above.

4 Illustrating approaches to internal assessment

4.1 We are not short of examples of implementations of internal assessment for summative purposes in the last 50 years; it would be difficult to describe more than a few of them, and there have been some descriptions of some aspects incorporated into the main discussion in this review. Rather than attempt to describe distinct models or select full examples (which would need a great deal of detailed description) I have extracted a number of illustrations of specific aspects of the conduct of internal assessment, within the taxonomy in the previous section, discussing their key features, effects and outcomes. Each illustration can be referenced to the four main elements discussed in the previous section.

Illustration 1: an A level coursework unit on the study of enzymes

This unit formed part of the 1993 version of a Biotechnology module within the Wessex A level Project.

All modules within all subjects in the Wessex project were constructed around the use of 5 skills and the activities in this unit related to these skills, which were

Seek out and retrieve information; Design and plan an investigation; Carry out an investigation; Interpret and draw conclusions; Communicate findings

This is an example of an optional activity on fermentation. It originates externally and lasts for 3 hours. Only the outline of the problem is provided although the student is provided with some background information on industrial fermentation processes, as part of a module booklet that explains the basis for all the work in the module. The activity is to be assessed against all 5 skills, the criteria for which are general and not task-specific, and included in a teachers' guide. Assessment is by observation of the student's work and of the content of the report. Quality assurance is through moderator visits, but these may not cover work on every module.

Fermentation practical

Use a commercial mini-fermenter (or design your own).

Follow yeast growth aerobically or anaerobically (the former requires and air pump) over a period of 48 hours, taking samples at suitable intervals. Growth of the yeast can be checked by (a) optical density (% transmission in a colorimeter) or by (b) cell count (haemocytometer). The yeast needs to be grown in yeast broth with 10% glucose or sucrose.

A growth curve can be plotted: (a) % transmission against time or (b) cell count against time. It may also be interesting to include a plot of (c) pH against time (if you have a pH probe) (d) temperature against time (if your fermenter is not thermostatically controlled).

In practice, work of this kind usually took more than the allotted time and some students needed a considerable degree of support. Some students felt that this was a difficult way to get an A level. Moderators were only able to sample reports of practical work and had to infer assessments relating to the conduct of the study. On the other hand, work of this kind demanded a considerable amount of innovation and decision-making by students whose skills were highly regarded.

Illustration 2: coursework in the Schools Council History 13-16 project

The following is taken form the 1979 moderator's report for this joint CSE and GCE O level examination.

Students were required to undertake coursework that formed 40% of the total assessment. They were to do between 4 and 7 written assignments totalling no more than 5000 words, covering all three of the units shown in the table below. Moderation was through a network of visiting moderators, each working with 2 schools, and having a specific staff development role as well as a quality control one.

General objectives were specified for each unit, as follows.

UNIT	GENERAL OBJECTIVES
Modern World Studies (10% of exam total; at least 1 assignment)	Analysis of CAUSATION and MOTIVATION
	Interpretation of the CURRENT situation in the context of PAST events
Enquiry in Depth (10% of exam total; at least 1 assignment)	EMPATHETIC RECONSTRUCTION on the ways of thinking, ideas, attitudes and beliefs, characteristic of people of a different time and place
	Analysis of the role of the INDIVIDUAL in history
History Around Us (20% of exam total; at least 2 assignments)	PERSONAL investigation and description of a site
	Correlation of ARTEFACT and DOCUMENTARY sources
	Relation of a site to its HISTORICAL CONTEXT

These objectives permeated the whole course and were discussed extensively in general support materials and development programmes. For the coursework teachers were told that, within the requirements specified in the table and within broad and specified limits, the weighting to be attached to each objective were flexible and at their discretion. There were, additionally, 3 optional objectives that could be assessed and teachers were able to specify others, but their proposals had to be justified to their moderator.

Teachers were provided with indications of the range of permissible assignments and earlier work done by students was used as a basis for staff development activities of an agreement trialling type (that is, discussion of the assessment criteria to be used and the applicability of these to a range of examples). The chief moderator also provided a comprehensive written commentary (almost 100 pages of typescript in 1979) on the coursework in each year of the examination.

The following is an extract from the chief moderator's report, illustrating some of the strengths and limitations of the coursework submitted. It relates to the History Around Us unit.

> Though there was much fine work done by many pupils, showing a great deal of planning, preparation, interest and hard work, it was sadly true that many teachers seemed to have no idea of the objectives of this section, and marks were, for example, awarded for descriptions of sites which were no more than a couple of sides of un-illustrated writing, showing little evidence of personal investigation. Many of the assignments were like traditional Local History projects, containing, for example, a general description and pictures of a town, and not really meeting the desired criteria at all.
>
> 1 Site description
>
> Some schools produced interesting, well-planned, well-illustrated work, based directly on guided observation, and on a subject which was intrinsically rewarding from many angles. Others chose one or two historical buildings, not necessarily related except by location, with a description, often based on the guide book…It is not the length of the work which matters so much as the quality of planning and observation and the inclusion of relevant illustrative material, all of which should show the mark of the pupil's own investigations, however simple these may be.
>
> 2 Historical background
>
> This was too often interpreted to mean class notes on 'Roman Britain' or 'Castles'. Unless the work is tied in closely with the chosen site(s), it loses much of its purpose. It should not be merely a chronological trot through *some* related (or in *some* cases, unrelated) topic, neither should it be merely a list of facts and masses of hand-outs. The background is intended to link the site chosen both to its historical context, specifically, and to other sites. ….
>
> 3 Use of sources
>
> … The whole question of assessment done in schools needs some attention, since even schools which had planned and structured the work with great thought and care did not always follow the criteria for assessment adequately. It was sometimes difficult to see even how marks had been gained in this section, since there was no obvious use of other source materials.
>
> 4 Imaginative reconstruction
>
> …It is very hard to strike a balance between an historical, factual account and an imaginative essay relying mainly on fantasy. Perhaps candidates could be encouraged to ask more questions of their material and their chosen site e.g. 'How would I have felt under such specific circumstances, remembering the different attitudes of the time?'

Illustration 3: an NVQ assignment specification

This is an example of an assignment specification for a level 3 NVQ and covering some aspects of key skills. It is expressed in quite general terms (almost as a series of notes) and does not specify the performance criteria that will be addressed, although candidates at this level could be expected to be familiar with the requirements of the NVQ and, probably, those of the key skills.

As part of your level 3 Admin qualification it would be useful for you to undertake a project based on the company you work for. If you do this thoroughly enough it can be used to cover many criteria from several units especially unit 5. If you are on the Modern Apprenticeship scheme it will have the added advantage of helping you work towards your extra key skills as you are specifically required to learn all you can about your company.

Compile this project separately using as many different kinds of evidence from as many sources as you can. For instance typed text, databases, spreadsheets, brochures, extracts from books, maps/plans, pictures, photos, lists, correspondence/replies, memos, e-mails etc. It should have an index and a bibliography showing where you obtained your information. It can include questionnaires, surveys etc.

An action plan showing how you intend to go about collecting this information would be very useful also; it need only be a fairly rough guide.

Your project could be something like this:

Background/history of company – find out from staff members, brochures and library. Photos and plans could also be included.

List of products/services supplied by your company and any associated information - this can be presented as a database or table or even spreadsheet/graph. Could you write away for any information from your suppliers? Any letters composed by yourself, and their replies, would be good.

Competitors – who are they, how do they differ from your company? Send away for their literature.

Marketing – how your product is marketed, show advertisements, articles, brochures etc.

Staff members – position/role in company, photos? database? Devise a useful questionnaire for the staff on health and safety for example; the results could be presented as a pie chart maybe.

Your job – a typical day, or a flow chart showing how you fit into the ordering process through to installing the end product.

Training and Development – have you been on any training courses through your company? Are there any set policies, have any other members of staff been on courses, were they good, where were they held? Would you like to go on any? Are there any ways you could develop your own position?

Although this may appear very general, even vague, it will operate in the context of a dialogue between the assessor and the candidate, be related to other work that will be included in the portfolio, and be capable of revision and extension in the light of discussion. On the assumption that decisions about the scope and form of the work and its investigationary part are done by the candidate, this is a good example of a relatively open-ended requirement that can be met in a workplace setting of the candidate's choosing. Because the work will be paper-based it is capable of being viewed by both internal and external verifiers. We can be fairly confident (because this is typical of many NVQ assignments) that the assessor and verifiers would be comfortable with this type of approach and regard it as a valid and reliable method of assessing certain aspects of competence whilst maintaining a degree of candidate control over the work.

This example appeared in a collection of portfolio exemplar materials issues by QCA in the late 1990s in support of the key skills pilot. It was accompanied by a couple of examples of candidate responses.

Illustration 4: some criteria for assignment writing

Early in the development of GNVQ a review entitled A Year in GNVQ 1992/3 offered the following requirements for good assignment writing.

Planning an assignment programme

An assignment programme is likely to

- be assembled as a team activity
- be seen as an integrated whole
- result from a mapping of the requirements of the vocational and core skill units into a coherent learning programme.

Describing assignments

Each assignment is a learning structure which enables students to

- work towards the fulfilment of one or, preferably, more than one of the performance criteria
- generate evidence of their learning, and of their accumulated attainments.

One of the principal tasks of the GNVQ team is to design assignments which

- allow students to meet the criteria
- are able to accommodate individuals' learning requirements
- are flexible enough to accommodate a range of prior experiences.

Making assignments

Assignments are likely to

- be neither excessively long or short
  - if very long students' learning may become narrowly focused, and lack of success or progress may become reinforced
  - if very short the total learning experience may be atomised
- contain sufficient support for student learning, but not so prescriptive as to prevent individual initiative, and management of own learning
- have a capacity to meet a range of individual needs
- be structured so as to allow evidence of learning to emerge
- be accompanied by mechanisms for planning, evaluation, assessment and recording.

An assignment programme will allow for a wide range of learning strategies, generating a range of evidence. Individual assignments will typically be accompanied by handouts describing the activities which may be undertaken or the tasks involved, and by other materials (describing methodologies, approaches, materials, literature and so on).

There is a wide range of approaches to the construction of assignments. Some care is needed to see that what is generated is not over-prescriptive, so that it prevents students generating ideas and approaches for themselves, as they develop the capacity to manage their own learning.

In practice, assignments have tended to become more prescriptive, and be presented as a series of tasks, linked to particular aspects of the specification; this is particularly the case at lower levels where a whole class group may do identical assignments, for insertion into portfolios. Assessment methods have tended to focus very strongly on the generation of paper-based products.

Illustration 5: A structure for assessment in a care NVQ

A great deal of assessment in NVQs and some in key skills uses observational methods. These require some organisation of opportunities for assessors to see candidates working on normal workplace tasks. This is an example of a planning sheet used in care (it refers to standards in use in the mid-1990s).

NVQs IN CARE
PLANNING SHEET

Planning for: Enable clients to maintain personal cleanliness

Unit: 29 Element: a and b

Performance evidence: a) 3 clients & problems
b) 3 clients & sensory equipment

RANGE	Client 1 AA	Client 2 MJ	Client 3 NP	Client 4 DW	Client 5
Bath etc	✔ Wash	✔ Bath		✔ Bed bath	
Wash hair		✔			
Shaving	✔				
Nail Care			✔		
Oral Hygiene	✔	✔			
Soreness/ discomfort	✔				
Changes in condition		✔			
Factors affecting cleanliness				✔ Questioning	
Imminent factors				✔ Questioning	
Combing/ brushing hair	✔	✔			
Makeup		✔			
Dressing/ undressing	✔	✔ Leg support	✔ Hearing aid	✔ Hearing aid	

On the next page is an assessment planning sheet on which a similar set of activities is related to the performance criteria and the evidence gathering method used or to be used.

Illustration 5 cont

ASSESSMENT PLANNING

DATE PLANNED OR EVIDENCE OCCURRED	RANGE	PC TO BE MET	EVIDENCE GATHERING METHOD
8/4/95	AA - get up and dressed am. All over wash, shave, dentures, brush hair, dressing	29a: 1,2,5, 6,7,8,12, 13 29b: 1,2,6	4: explanation of process 8: questioning
	AA Dad (211)		9
10/4/95	MJ Bath 8-00am Bath, hair wash and brush, dressing, leg support, makeup Commode 211a,b,c Transferring 27 a & b	29a: 1,2,3, 7,8,9,12,13 29b: 1,2,3, 4,5,6	1: natural obs. 8: questioning
?	NP Nail care Hearing aid		1: natural observation
by end of May	Case study on personal hygiene of a Muslim lady (AWARENESS OF CULTURE)		Project

Preparation needed	How
- Read through personal hygiene procedures - Bed bath - Shaving	- Office/discussion me - work with Kay - demonstration/ discussion on care of equipment

Signed: Candidate _____ Assessor _____

Date _____ Date _____

Other records will relate to the outcomes of specific observations, written by either the candidate or the assessor, and witness statements, both of which will be incorporated into a candidate's portfolio. This documentation forms evidence that can easily be viewed by internal and external verifiers.

This is a tightly controlled structure that is likely to meet awarding body requirements, be very reliable (largely because of the large number of observations) but which, in the hands of an unimaginative or dominant assessor or when subject to very tight verifier control, can take almost all initiative away from the candidate.

Illustration 6: views of awarding body quality assurance and quality control procedures in academic qualifications

Ecclestone & Hall (1999) have provided a summary of views of a range of procedures used for quality assurance and control by awarding bodies.

	internal moderation	*moderation and standardisation meetings*	*exemplar materials*	*examiner reports*	*postal moderation*	*subject meetings*	*awarding body guidance*	*official syllabus*
Teachers	Seem to adapt internal standardisation to be developmental as well as to meet AB requirements Not clear how far staff separate explicitly informal checking of uncertain grades and formal exercises to standardise a sample Essential part of QC but institut-ional resource pressures reduce time on these exercises	Seen as useful although more so for new staff as a way of learning the standards Resource pressures (time to attend, number of opportunities offered by the AB) mean that attendance is sporadic and ad hoc Being an AB moderator is seen as very developmental and useful	Seen as useful and important Variation in how far teachers use exemplars with students a) as revision b) as general way of inducting students into required standards	Seen as useful and important Variation in how far institutional managers use these and/or follow up issues	Clearly understood as a QC procedure An established part of QC (compare with vocational teachers)	Seen as useful and important but opportuni-ties are a) patchily distributed amongst teams b) sporadically offered by awarding bodies Complaints that they tend to be dominated by the latest statutory requirements rather than subject expertise	Seen as useful but patchily distributed within institutions Most staff adapt criteria specifications for students to use	Largely seen as given and not negotiable Teachers adapt as much as possible to reconcile national standards and local conditions of student motivation etc.
Awarding body officers	See this as a formal QC procedure, although they do not use this language Variation in language between standardisation and moderation Guidance on procedures to follow but not prescriptive	External moderation exercises generate materials for teachers to use in internal moderation Results and exemplars from standardisation exercises are used to train moderators	Exemplars are provided for teachers to use in internal moderation	ABs have 2 types of report: coursework reports about annual standards and a centre's own performance and subject reports about issues Reports are used internally to train moderators and review QA/QC procedures	Formal QC procedure to moderate a sample that has been internally moderated in a centre beforehand	Recognition of value and need for more opportunities for teachers Problems of resources both from AB and centres themselves	Some variation in how ABs distribute guidance to centres ABs do not see their role to provide generic guidance about assessment Recognition that ABs are being required to take a more regul-atory role e.g. auditing, monitoring standards etc.	
FEFC inspectors	Seen as essential for QA and QC but affected by resource issues See manage-ment support for internal moderation as essential See a need for compliance from managers and staff to QA and QC: weak-nesses they see when compliance is not total	Recognise resource pressures and constraints		Seen as part of holistic approach that an institution must incorporate into QA and QC			View that guidance is sporadically used and distributed	

Illustration 7: using agreement trials

The following is an extract from an awarding body procedure document produced in the late 1980s in connection with the introduction of GCSE.

> In agreement trials a consensus marking standard is established as a result of a series of meetings held on a local or regional basis embracing all the teachers who are involved in the assessments to be moderated. To be of maximum benefit these agreement trials should be held shortly before the assessments are made. For many subjects, special prepared video presentations are an essential means of presenting the Group's standards. A moderator appointed by the Group should be present to lead discussions and ensure standards. The purpose of the trials is to ensure that all teachers are capable of identifying the appropriate standards of work required. A selection of candidates' work covering all the grades to be awarded, either from a previous examination or from the present examination, is taken to the meeting. Agreement is reached by consensus on the mark to be awarded in each case, and the reason for it. The participants later assess their own candidates in the light of discussion at the agreement trial.
>
> Moderation by agreement trial has the advantage that attempts are made to standardise marking before the teacher makes his or her assessment. However, it requires a considerable organisation at the local level and is expensive of teachers' time. It is expensive to operate, particularly for examination whose entry is thinly scattered over a wide geographical area.

There has traditionally been a distinction between agreement trialling and consensus moderation; the former is a professional development process by which teachers and assessors use evidence generated by students as a basis for clarifying (but not usually generating) assessment criteria, and identifying critical decisions that enable them to agree upon a standard. Some of the processes advocated in the mid to late 1990s, where teachers in a school department collected, and analysed and stored examples of students' work in a portfolio, to provide a reference point for agreed interpretations of the assessment criteria, was essentially an agreement trial process.

For a much more detailed discussion of agreement trialling see Wilmut (1997).

Consensus moderation has generally been seen as a post-assessment process of agreeing final marking or grading decisions. In it, teachers bring samples of work done by their own students and which are re-assessed by others, with discussion and review being used to reconcile disagreements to the point where consensus is reached on each of the items. Such meetings are usually conducted by awarding body moderators who are expected to guide and manage the review process and, ultimately, decide on where a sufficient degree of agreement exists. Where this is not achieved the usual fall-back is for the moderator to make an assessment and impose it.

Both processes are very time and labour-intensive, and it became increasingly difficult in the 1990s for awarding bodies to meet the costs involved and for teachers to be made available for participation in such meetings. The argument that teachers gained a great deal of experience from such meetings has some force, but this does not always appear to have been disseminated through departments in schools and may be a quite inefficient way of conducting professional development. In some case, consensus moderation could become rather casual, with untested assumptions being made about teachers who always got the standard right and others who did not.

Agreement trialling and consensus moderation have largely disappeared from the conduct of internal assessment, to be replaced by verification and moderation by individuals, both internally (particularly in the case of vocational and occupational qualifications) and externally. Although the decisions are based on an individual's judgement, most teachers and assessors tend to develop a good relationship with their visiting moderator or verifier and use the occasion to glean information and test their understanding of the requirements and standards for the internal assessment. Many teachers and assessors would regard postal moderation as the greatly inferior system, since it provides no such direct contact.

Some moderation systems have been built around local consortia, particularly in the days when LEA advisers provided a focus. This system was briefly used for key stage 1 of national curriculum assessment and is not to be re-introduced. However, centre-to-centre contacts through such consortia are much weaker than they once were, and may provide a less effective environment for moderation than was once the case.

Illustration 8: a minor and tightly controlled coursework component of a summative assessment

The illustration below is not a specific example, but is a typical internally assessed component that is operated so as to meet a specific validity requirement, ensure the highest achievable reliability, and be economical to operate.

Rationale	Internal assessment is introduced into a previously external assessment on the grounds that there were some domains that were not being adequately assessed in the written examinations. The weight given to the internal assessment is lower than that given to the external, and its conduct is tightly controlled.
Typified by	A wide range of coursework implementations in all public examination provision since 1950. Coursework has taken a number of forms, such as a project, design, construction or investigation; a performance; a practical test to be conducted internally. In some respects it also covers some externally set assignments, although in some vocational settings these carried more weight than external tests.
Task choice	Task origin: the awarding body generates a set task. A slightly more flexible arrangement offers a small choice of tasks. Tasks may be retained over several years. Task duration and importance: the work may occupy up to, say, 30 hours of time within the learning programme; it is normally positioned quite late in the learning programme.
Task prescription	Prescription type and detail: a set of generic but quite specific performance or outcome related criteria, or (when there is a single task) criteria linked specifically to the task or elements of the task.
Authentication	Centres are expected to warrant the authenticity of students' work.
Assessment	Assessment role: usually additive and compensatory. Assessment judgement: results in a mark; marks are later added to give a total. Assessment occasions and methods: one occasion, either at the end of the work, based on an outcome (report, artefact etc) or performance; this therefore seen to be coursework rather than continuous assessment.
Quality Assurance	Centres are expected to undertake internal quality assurance so that the awarding body may treat the centre as a whole and not on a class-by-class basis. A method of moderation at a distance (postal moderation) will be the most common method. This will entail the re-marking of a sample of work by an external moderator; the sampling method will be specified by the awarding body. The normal assumption is that, unless there is very strong evidence to suggest erratic internal marking, the rank order of students will be preserved and adjustments will be made to the mean and standard deviation of the internal marks, regressed on those of the moderator. Where this cannot be done, all the work is called in and re-marked, but this is an expensive and time-consuming process that cannot be done too much or too often. Where a performance element is the only output from the internal assessment visiting moderation (possibly again seeing a sample only) is the preferred alternative. Systems that have operated such moderation within local networks of centres have also been used.

This approach to coursework is proven and manageable on a large scale. Its disadvantages lie in the very limited scope for choice or decision-making by teacher or student; the production, by the student at the teacher's prompting, of standardised responses; and the limited connection with the learning programme. Moderation methods provide limited input to learning. In these respects, however, this type of coursework component is no worse than the written examination, though it is little better either.

5 A rationale for internal assessment

Justifying assessment by teachers

5.1 The earlier review (in Section 2) of the growth of teacher assessment for summative purposes identifies some variety in the justifications for its growth. The views that were expressed can be summarised as

- teacher assessment should be included in examinations because this enables objectives to be assessed that were inaccessible to written examinations
- there was a capacity to bring teacher assessment closer to classroom assessment practice
- negative backwash of summative assessment on the curriculum would be reduced and a closer link between learning and summative assessment established
- the values and pedagogic requirements of teachers could be better met if they had some degree of control over the summative assessment
- it was appropriate that responsibility for summative assessment should be shared between teachers and examinations boards and partnerships of this kind were beneficial
- the support and control exercised by examination boards would enhance teachers' assessment expertise
- internal assessment was the only appropriate way to determine learner competence in workplace and other vocationally-related settings.

This probably is not an exhaustive list but it does form the basis for most of the discussion of issues that follows in this and later sections.

5.2 Three important strands seem to emerge from this list. The primary strand is concerned with policy and control: where does the decision-making and responsibility lie for the conduct and outcomes of the assessment? Discussions about the other two strands seem incapable of resolution unless there is some clarity in principle about where this control lies and this is discussed in the remainder of this section. The second strand is the relationship between assessment for summative purposes and assessment that is primarily for learning in the classroom; this is the subject of Section 6. The third strand is concerned with the validity and reliability of the summative assessment in relation to the purposes for which it is needed; I have examined this in some greater detail in sections 7 and 8. The first I will deal with briefly here, since it impacts on the perceived role of the teacher as an assessor and the roles of examining boards or awarding bodies, regulators and the government.

The teacher as an assessor

5.3 A discussion of the roles and responsibilities of the teacher as an assessor for summative purposes will relate to the involvement of other stakeholders: awarding bodies, regulatory authorities, a range of user stakeholders, the general public and the government. The issues become part of the debate on 'standards' in which what is known about examinations and their results are used as a core part of the discussion of the curriculum and students' performances in it, whether these are rising or falling and

whether current provision is adequate for the needs of the 21st century (MORI/CDELL, 2002).

5.4 Cohen and Deale's view of the approaches to internal assessment was cited in para 2.11 as one part of a wider and ongoing debate about the role of the teacher as assessor. Another common view some years ago was that the adoption of teacher assessment for summative purposes simply involved fine-tuning or adapting assessment practices which were already part of the teachers' professional activities. That now seems rather simplistic but there is still a hope that the workloads for both teacher and learner associated with assessment for summative purposes might be reduced if this were somehow to be conflated with normal classroom assessment activities.

5.5 Certainly, in the 1980s, many teachers chose to be involved in developing approaches to internal assessment within the provisions of the new GCSE general criteria because they wanted to develop new approaches to assessing their subject, or to be able to assess more flexibly (Torrance, 1995). This would involve the introduction of more practical work in science or more local studies in geography, with the capacity to assess such work *in situ* and to provide students with shorter-term goals and a more continuous approach to assessment. The essential point, as had earlier been the case with CSE, was that many teachers were engaged with this for curriculum reasons rather than primarily as a process of assessment reform, a motive that was also a feature of the involvement of some in the development of graded tests and in records of achievement. In this sense teacher assessment was teacher-driven and articulated in relation to specific pedagogic demands.

5.6 Torrance suggests that where teachers have become involved in the use of examination coursework that has been imposed from outside (perhaps without a clear pedagogic rationale) their attitudes have been less clear and often less supportive. Rationales that connect external control over coursework with better control over the validity and reliability of the assessment have carried less weight and many teachers have quickly become concerned about the additional workloads involved and about the technical details of the conduct of the assessment processes. Some commentators (such as Black, 2002) have doubted the relevance of such externally controlled internal assessment to students' learning development. This emphasis has been reinforced by some of the training and support provided by awarding bodies and others that has, not unreasonably, been less concerned with developing teachers' generic assessment skills than with the mechanics of the conduct of a specific qualification and its quality control, perhaps springing from a concern to achieve the highest possible reliability coupled with the minimum demands on teacher time. There are interesting parallels here with other initiatives. For example, support for the introduction of Records of Achievement often focused on the mechanics of the process and the effects on students, but did not always help teachers to provide the broad intentions of the programme and enable them to work through the theoretical and practical problems of implementation (Broadfoot *et al*, 1988).

5.7 Robbins (1998) also identified this tension and noted that the purposes for internal assessment are

- not static, and
- not always shared between teachers and the awarding body.

He wanted to see the purposes for internal assessment as the primary and central issues, such that matters of validity, reliability, manageability and the like all stem from purposes. Evidence from work on examination grading over a long period (Christie & Forrest, 1981; Cresswell, 1996) provide ample support for the view that perceptions of the purposes of coursework need continual re-defining for various groups: the public, teachers, students, boards, and so on. Moreover, the values made explicit in a syllabus are not necessarily those that underpin learning in the classroom and the formative assessment used there. However, involving teachers in the development of shared common purposes for coursework is difficult and costly in large-scale public examinations.

5.8 Teachers might justifiably feel that their control over assessment had been steadily eroded in the last 20 years. They might point to the GCSE criteria of the late 1980s that were seen to lack flexibility, to apparently arbitrary decisions to limit the use of coursework, to the dominant status of tests and the increased paperwork that accompanied teacher assessment within the national curriculum and to the increase in pressure to ensure that their assessments could be defended against criticism. They would probably have to accept, however, that the quality of summative assessment has improved considerably over the last 20 years and that far more students are gaining good results, that support materials for teachers and students are more readily available in far greater quantity and that their professional skills in assessment have improved.

5.9 We might ask whether teachers would sign up to internal assessment for summative purposes that, in relation to issues of control,

- might reduce the role and workload of the terminal examination
- further engaged them in the summative assessment process and enhanced their professional skills
- enabled them to relate aspects of summative assessment directly to local requirements, so that performance might be assessed in naturalistic settings and contexts, better meeting the learning demands of their subjects.

They might then also want to endorse a pattern of assessment that had a greater proportion of teacher assessment and that would therefore, as a whole,

- have extended validity and curricular legitimacy
- offer the opportunity to create a better symbiosis between formative and summative purposes
- offer the possibility of enhancing learner control and the possible development of learning skills.

They might recognise that there may be some consequential effects, such as

- a shifting of workloads onto the teacher
- a transfer of pressures from the examination onto work done throughout a programme
- a reduction in the reliability of the whole assessment
- a greater investment in development and training at a local level

- a public perception that there is a dilution of standards that are seen to be vested in the externality and commonality of external examinations and tests.

5.10 The taxonomy of Section 3 suggests that there are several areas of control in each of which there are balances to be struck between teachers and others stakeholders. Closely specified tasks that are to be assessed by closely specified criteria will leave teachers with little autonomy in the conduct of internal assessment and little opportunity to establish linkages with formative assessment. A more flexible provision that offered more control to teachers might attempt to establish frameworks or criteria for the choice of tasks and generic criteria for assessment, but this may involve accepting lower levels of validity and reliability for the assessment. Systems of quality assurance and control may be managed entirely by the awarding body or may involve some degree of internal monitoring of the assessment process. There do not seem to have ever been rational and systematic debates of these factors and it is hard to see how progress can be made without it.

5.11 In the process a number of beliefs may need to be identified and critically examined, perhaps with the aid of some systematic research. It is, for example, difficult to see how internal assessment for summative purposes can be expanded in an environment where it is widely regarded as unreliable and vulnerable to bias. Perhaps the difficulties with reliability and bias may need to be more accurately placed in the context of the reliability and bias in all forms of assessment and the reliability and bias associated with different approaches to internal assessment examined more exhaustively so that appropriate choices of method can be made or patterns of support developed that will effectively reduce the problems.

Learning from the national curriculum

5.12 There has been a rather rigorous examination of some of these propositions in a debate between Newton (2003a, 2003b) and Wiliam (2003) in relation to national curriculum teacher assessment (but with resonances to assessment for qualifications, though there are arguably higher stakes to contend with there). Some of this is concerned about the evidence available for making assertions about what will and will not work by way of establishing a relationship between teacher assessment for formative and summative purposes. But a couple of important issues emerge, particularly in relation to the practicalities and policy implications of seeking changes in the present arrangements.

5.13 First is whether any mechanism for the creation of summative assessments from some set of formative assessments made by teachers will overcome problems of workload and problems of validity and reliability at levels required for summative reporting purposes. Newton suggests that both issues are problematic and further, that there is insufficient replicability in culling teachers assessments across many domains to generate the required levels of reliability for, say, GCSE purposes. Moreover, bias may enter the assessments in ways that cannot be detected or controlled.

5.14 Newton is cautious in accepting the proposition that external summative assessments narrow the curriculum, although he is referring only to national curriculum tests rather than to externally-specified internal assessment requirements and does appear to regard the need for accountability as the more restricting influence. Here, the discussion of national curriculum assessment appears to become divorced from that related to 14-19 qualifications – not only are the stakes higher in the latter case but the control over teacher assessment is quite different.

Summary

5.15 In relation to the purposes underlying teacher assessment for summative purposes:

- A range of justifications has been offered including that it enables objectives to be assessed that are inaccessible to written examinations; it brings teacher assessment closer to classroom assessment practice; negative backwash on the curriculum is reduced; a closer link between learning and summative assessment is established; the values and pedagogic requirements of teachers are better met if they have some control over the summative assessment; partnerships between teachers and examination boards are beneficial; teachers' assessment expertise is enhanced. (para 5.1)
- It is said that internal assessment is the only appropriate way to determine learner competence in workplace and other vocationally-related settings. (para 5.1)
- The primary issue in discussing purpose is policy and control: where should the decision-making and responsibility lie for the conduct and outcomes of the teacher assessment? Discussions about other issues (such as about the relationship between assessment for formative and summative purposes, about validity and reliability and about quality control) are incapable of resolution unless there is some clarity in principle about this. (paras 5.2 & 5.7)
- The purposes for internal assessment are not static, and not always shared between teachers and the awarding body. However, involving teachers in the development of shared common purposes for coursework is difficult and costly in large-scale public examinations. (para 5.7)
- Perceptions of the purposes of coursework need continual re-defining for various stakeholder groups. (para 5.7)
- The values made explicit in a syllabus are not necessarily those that underpin learning in the classroom and the formative assessment used there. (para 5.7)

5.16 In relation to the engagement of teachers in summative assessment:

- In the past many teachers were engaged with summative assessment for curriculum reasons rather than primarily as a process of assessment reform. In some situations teacher assessment was teacher-driven and articulated in relation to specific pedagogic demands. (para 5.5)
- Where teachers have become involved in the use of examination coursework that has been imposed from outside their attitudes have been less clear and often less supportive. Rationales that connect external control over coursework with better control over the validity and reliability of the assessment have carried less weight and many teachers have quickly become concerned about the additional workloads involved and about the technical details of the conduct of the assessment processes. (para 5.6)
- Training and support provided by awarding bodies and others has been less concerned with developing teachers' generic assessment skills than with the mechanics of the conduct of a specific qualification and its quality control. This has sprung from a concern to achieve the highest possible reliability coupled with the minimum demands on teacher time. (para 5.6)

- Balances need to be struck between teachers and others stakeholders in relation to the degree of control over the choice and specification of tasks, acceptable levels of validity and reliability and the relationship between internal and external control of quality. There do not seem to have ever been rational and systematic debates of these factors and it is hard to see how progress can be made without it. (para 5.10)

5.17 In relation to getting a better understanding:

- Reliability and bias in internal assessment should be more accurately placed in the context of the reliability and bias in all forms of assessment. The reliability and bias associated with different approaches to internal assessment should be studied more exhaustively so that appropriate choices of method can be made or patterns of support developed that will effectively reduce the problems. (para 5.11)
- There is a need to explore whether mechanisms for creating summative assessments from some set of formative assessments made by teachers will overcome problems of workload and problems of validity and reliability at levels required for summative reporting purposes. (para 5.13)
- Research is needed on whether there is insufficient replicability in culling teachers' existing assessments across many domains to generate the required levels of reliability for summative purposes and whether there is bias in the assessments that cannot be routinely detected or controlled. (para 5.13)
- There should be further exploration of the proposition that external summative assessments narrow the curriculum. (para 5.14)

6 Internal assessment for formative and summative purposes

Interactions between formative and summative assessment

6.1 The term 'formative assessment' has been in use for more than 30 years, but has recently been somewhat displaced by the more self-evident term 'assessment for learning' to distinguish it from summative assessment that may be seen to be 'assessment of learning'. There is a danger that any assessment done by a teacher or done in a classroom may be assumed to be assessment for learning, but this is clearly not the case, and the Assessment Reform Group and others have worked hard in the last few years to identify and publicise the characteristics of this type of assessment. There is also a danger that, having labelled internal assessment as assessment for learning it is then hi-jacked for summative purposes; we are beginning to understand some of the risks in doing this.

6.2 Stobart (2003) has described the position taken by the Assessment Reform Group regarding assessment for learning. He has provided the Group's definition of this form of assessment 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."

A basis for assessment for learning has been written into 10 research-based principles designed to guide classroom practice (Assessment Reform Group, 2002b).

6.3 For many years there has been seen to be a tension between the purposes of formative and summative assessment, and the view has been that it is better to keep them apart. Thus Harlen *et al* (1993) were of the view that formative assessment should be carried out "... with the rigour and reliability necessary to be effective in improving pupils' learning" and teachers need skills and materials to enable this, whilst recognising the dangers in using assessment for multiple purposes, making it important to maintain the distinction between formative purposes and summative purposes. They usefully divided summative assessment into processes of 'summing up' (creating a final record) and 'checking up' (establishing what has been attained at the end of a programme segment). A teacher may be involved in both at different times, but checking up may come very close to an assessment that is for learning. Gipps (1994) bore out the tensions that are involved:

> "Where school-based teacher assessment is to be used for summative purposes then the relationship between teacher and pupil can become strained: the teacher may be seen as judge rather than facilitator. This uneasy dual role for the teacher which ensues is a result of the formative/summative tension."

6.4 In recent years, teachers have drawn formative and summative purposes together, especially in their assessment for the national curriculum. They have, for example, culled test materials that have been designed for summative purposes and used them within their schemes of work as a way of monitoring progress. Students' need to generate evidence and have it assessed in relation to competence statements in a GNVQ or NVQ makes it very difficult to draw a line between formative and summative purposes so that Ecclestone & Hall (2000) comment that

> "…teachers tend to adapt their uses of internal assessment for their students and conflate formative and summative functions of internal assessment".

The resulting problem is observed by Sadler (1989):

> "Continuous assessment cannot, however, function formatively when it is *cumulative*, that is, when each attempt or piece of work submitted is scored and the scores are added together at the end of the course. This practice tends to produce in students the mindset that if a piece of work does not contribute towards the total, it is not worth doing."

Black and Wiliam (1998) note that a

> " ... tension between formative and summative assessment arises when teachers are responsible for both functions: there has been a debate between those who draw attention to the difficulties in combining the two roles... and those who argue that it can be done and indeed must be done to escape the dominance of external summative testing..."

6.5 If we are to establish a working relationship between the formative and summative roles we will need to arrive at a better understanding of the characteristics that make assessment an essential component of the learning process and how some of its features may be shared with summative assessment and some not. It will not simply be the nature of the instrument used; clearly, we cannot just say that any test or instrument sourced externally will automatically be summative and that any piece of less formal assessment done in a classroom will automatically be formative or for learning. Nor can we assume that an instrument developed for external summative purposes and used uncritically or inexpertly will satisfactorily perform a useful formative role; this may be a difficulty that current developments of all-purpose question banks in basic and key skills, made available for teachers' use, may compound. A resolution of the tension between formative and summative assessment will require teacher expertise in assessment that is concerned with principle and purpose and not just with process.

6.6 The discussion of the relationship between formative and summative purposes and their relationship within national curriculum assessment has been extensively reviewed in a range of key papers (e.g. Wiliam & Black, 1996; Black & Wiliam, 1998; Harlen & James, 1997). These include the influential discussion *Inside the Black Box* that appeared as a result of the Black and Wiliam review of formative assessment and that forms a key part of the current work of the Assessment Reform Group[3]. In this context, Black (1990) attempted to grasp the problems of the multiple uses to which assessment and the information derived from it is put; these have generally been held to be formative, for certification, for selection and for accountability. If the demand for accountability is legitimate (he claimed that it is now an inescapable constraint) it creates additional tensions between formative and summative assessment. These should be reduced as far as possible since we cannot have good formative assessment in situations where teachers' efforts are not valued, respected and trusted. There are some priorities: we should get formative processes right first since it is a prerequisite for both good learning and good summative assessment. Some headway can be made by linking, as far as possible, external instruments and methods to those being used by teachers, and by involving teachers in external processes and Black and Wiliam (1998) have

[3] See the Assessment Reform Group website at www.assessment-reform-group.org.uk

called for a greater investment in support for teachers making assessments, and in research and development. These views were reinforced in a more recent review (Black & Wiliam, 2003) in which they acknowledge

> "… that teachers need time, freedom and support from colleagues in order to reflect critically upon and to develop their practice, whilst offering also practical strategies and techniques about how to begin the process. By themselves, however, these are not enough. Teachers also need concrete ideas about directions in which they can productively take their practice and thus there is a need for work on the professional development of teachers to pay specific attention to subject-specific dimensions of teacher learning."

6.7 We should not under-estimate the levels of professional expertise in assessment required of teachers if they are to become key players in the resolution of the tension between formative and summative purposes for assessment rather than just implementing externally controlled initiatives. Black and Wiliam offer evidence that teachers become confused when assessment policies are changed; the pace of change in schools is often slow since teachers find it very difficult to change practices which are deeply embedded in their whole pattern of pedagogy. They cite work by Radnor (1994) who reported a lack of change in assessment agendas, despite a training programme within a project. Changing demands on teachers to act as assessors in relation to qualifications may require major shifts in pedagogical approach similar to those that accompanied the introduction of GNVQ or the national curriculum.

6.8 Radnor's work is also of interest because it offers some insights into the mechanisms of the assessment process. She reports a project in arts education, working with students in years 7-9 across several schools. Using agreed structures teachers in these schools made qualitative assessments through talk with pupils, undertaken as formative assessment, but in the context of the national curriculum. Three main issues emerged from the analysis of the records of these conversations.

- The first was concerned with teachers' perceptions of the student-teacher relationship. The intention was that the talk would allow an open response to the students' work and their attitudes to it, explored in all of its aspects. But the teacher's position as expert and as authority got in the way, and students often told the teachers what they thought they wanted to hear, and withheld some information and perceptions.
- The second was concerned with teachers' attitudes towards their roles as assessors. They found it difficult to break out of a mode of assessment that is concerned with making summative judgements. Radnor claims (probably correctly) that this is a consequence of the need to operate national curriculum assessment.
- Finally, teachers' understanding of what counts as valid evidence became the framework for the knowledge that students transmitted to them. Students were not telling teachers what they actually knew, but what was seen to fit into the frameworks as they understood them.

6.9 Radnor also suggested that teachers have a stake in the future of the school, as an institution, and therefore conducted assessment in a way that causes minimum disturbance for themselves and for the institution. It was not actually in their own interests to extend the assessment frameworks beyond what was needed by the institution. The difficulty, of course, is that the perceptions of the learning of the

individual, and the decisions about the next steps, are limited within pre-determined frameworks, and not operated within the structure of the students' knowledge; assessment is not truly 'for learning'.

6.10 Pryor & Torrance (1996) pursue similar issues, suggesting that there are two conceptually different processes at work within teacher assessment: these are 'convergent' and 'divergent', and they have worked with these approaches with teachers and children at key stage 1.

- Convergent assessment seeks to find out if a child knows a predetermined theory, is characterised by checklists and predetermined questions, curriculum-based and supported by detailed planning, and designed to check whether the next pre-determined stages of learning can start.
- Divergent assessment emphasises learner's understanding and seeks to discover what the child knows, is conducted as a joint pupil-teacher activity and uses open forms of recording. It is seen to strengthen classroom practice.

The requirements of convergent and divergent assessment both have a legitimate place in schools: "... what seems to be required is an approach to assessment that enables both convergent and divergent teacher assessment to be pursued at appropriate times."

Assessment and motivation

6.11 In the course of their paper Pryor and Torrance discussed the issue of child motivation, through assessment. The school in which they were working had a reward system that included the use of stickers, smiley faces and merit marks. Whilst such a behaviourist system may strongly support convergent approaches it may not allow insights into the full range of children's knowledge in a particular area. They cite the earlier work of Dweck (1989) and Lepper & Hodell (1989) which had suggested that "extensive reward systems might harm the motivation of children and contribute to their adopting performance rather than learning goals".

6.12 In what ways does assessment motivate learning? Harlen and Deakin-Crick (2003) addressed this issue head-on through an EPPI study of the effects on students' motivation for learning of testing and other forms of summative assessment. The findings (summarised in a pamphlet issued by the Assessment Reform Group, 2002a) were based on the results on 19 research studies that met a range of criteria (relating to relevance, methodology, quality of evidence, dependability, etc.) and relate only to compulsory education. They explore issues of motivation using a model that distinguished extrinsic from intrinsic motivation. Although the review was designed to relate specifically to the effects of summative assessment, the conclusions do tell us a lot about the differences between this and assessment for learning and how the negative effects of external testing may be minimised and its positive impacts utilised.

> "Many of the findings… suggest avoiding drill and practice for tests, de-emphasising tests by using a range of forms of classroom assessment and recognising the limitations of tests, preventing the content and methods of teaching from being limited by the form and content of tests and taking steps to prevent children being faced with tests in which they are unlikely to succeed. … However, rather than indicate only what should be avoided, there are more positive messages for action that teachers and schools can take to ensure that the benefits of

summative assessment can be had without negative impact on students' motivation for learning. The following were identified:

a. Promote and engage in professional development that emphasises learning goals and learner-centred teaching approaches to counteract the narrowing of the curriculum.

b. Share and emphasise with students' learning goals, not performance goals, and provide feedback to students in relation to these goals.

c. Develop and implement a school-wide policy that includes assessment both for learning (formative) and of learning (summative) and ensure that the purpose of all assessment is clear to all involved, including parents and students.

d. Develop students' understanding of the goals of their learning, the criteria by which is it assessed and their ability to assess their own work.

e. Implement strategies for encouraging self-regulation in learning and positive interpersonal relationships.

f. Avoid comparisons between students based on test results.

g. Present assessment realistically, as a process which is inherently imprecise and reflexive, with results that have to be regarded as tentative and indicative rather than definitive."

6.13 In a conclusion directed at policy makers they observe that

"… current high stakes testing is failing to provide valid information about students' attainment for a number of reasons. For example, the tests are too narrowly focused to provide information about students' attainment and the consequences of teaching to the tests mean that students may not in reality have the skills or understanding which the test is designed to assess, since teachers are driven by the high stakes to teach students how to pass tests even when they do not have these skills and understanding."

6.14 In 14-19 programmes there are a number of respects in which students' perceptions of assessments and their outcomes have changed radically over the last 20 years. It is a widely held view in schools and colleges that they have become very utilitarian in their views of what it is worthwhile to pursue; Ecclestone & Hall (1999) call this a "… strategic and cynical compliance with assessment requirements" that is something that teachers promote. It may not exclude enthusiasm for work, and may represent a commitment to the achievement of some long-term goal but where assessment systems are heavily standardised, students seem to be less engaged with the content and processes of courses. This is not an isolated piece of evidence (see below and in Section 9) and it suggests that we may need to construct and operate assessment in such a way that learners perceive that it is contributing to their wider progression goal, via a qualification. In a practical sense the distinction between formative and summative purposes almost disappears and the relationship between teacher and student becomes a 'professional' collaboration that generates its own motivation. The risk that it carries is the marginalisation of learning that is not seen to contribute to the gaining of the qualification.

Feedback

6.15 Formative assessment and assessment conducted by the teacher for summative purposes offer the possibility of giving students information about their learning, enabling them to make progress. But Sadler (1989) notes

> "... the common but puzzling observation that even when teachers provide students with valid and reliable judgements about the quality of their work, improvement does not necessarily follow."

6.16 Gipps *et al* (2000) suggest that there is evidence that feedback may inhibit improvement and, for it to be effective, Stobart (2003) has suggested that it should be clearly linked to the learning intention so that

- the student understands the success criteria and standard
- feedback focuses on the task and not the learner
- feedback gives cues about next steps
- feedback challenges, requires action and is achievable.

6.17 If feedback simply signals approval or disapproval, is only presented as composite marks or grades, or makes comments on effort rather than specific achievement in relation to shared criteria it may be at best ineffective and at worst counter-productive. Stobart suggested that improving the quality and extent of feedback may not be incompatible with the use of internal assessment for summative purposes and may become a feature of more sophisticated management of external assessment through the use of ICT. Of course, it will be difficult for this feedback to support learning if it is based solely on a single piece of coursework late in a programme; like the current script return facility, there is more potential value in feedback to teachers than to students who have already moved on.

6.18 Weeden and Winter (1999) reported attitudes to feedback amongst students at all stages of education from key stage 1 to post-16. The comments of post-16 students are particularly interesting: they reinforce the limited value of feedback that does not include comments and information related to the specifics of the work that they have done, but also emphasise their need to be able to take action that will bring particular pieces of work up to the required standard. This is especially relevant in GNVQ and similar programmes where a student may need to complete an assignment that meets one or more specific outcome statements, insert this in the portfolio, and then move on to the next stage. For such students feedback that is insufficiently focused is of little use.

6.19 There is a danger in a summative assessment environment that students' learning needs can become marginalised. In an extended study of occupational learning, and its transferability to workplace settings, Tolley *et al* (2003) looked in great detail at learning and assessment processes as they happened in real and simulated workplace settings. They used video evidence to probe the interactions between NVQ assessors and candidates and, in relation to several groups of trainees for different NVQs, they commented on the failure to identify learning and assessment opportunities and on the often limited feedback processes from assessors to candidates. They link this to candidates' capacities to manage their own learning; the following comment on hairdressing candidates illustrates the issues:

> "Candidates need… to develop as self-organised learners, because the onus will continue to fall on them to identify unplanned learning opportunities and to ensure that they receive feedback which illustrates the general lessons that can be learned from particular incidents. To that end they need to develop the ability to counteract their own marginalisation during problem solving episodes in the workplace. The evidence suggests that candidates are not acquiring the knowledge, understanding and skills which would help them to develop the 'independent capability' they need when confronted with new contexts, situations and contingencies."

6.20 It is possible that some forms of mastery assessment for summative purposes will not have the capacity to generate feedback that will improve learning. Simpson (1990) has described her work on the learning of biology students and suggested that criterion-referenced assessment only tells us whether they have learned as intended, not what they have actually learned (whether that be right or wrong). It does not, therefore, provide a basis for diagnosis and progression, limiting the resolution of the tension between formative and summative purposes.

Summary

6.21 In relation to the tension between formative and summative purposes:

- For many years there has been seen to be a tension between the purposes of formative and summative assessment, with the view that it is better to keep them apart. (para 6.3)
- There have been said to be tensions in the role of the teacher when involved in both formative and summative assessment. (para 6.3)
- More recently, teachers have drawn formative and summative purposes together by culling test materials that have been designed for summative purposes for monitoring and formative uses. It has been argued that formative and summative roles must be combined in order to escape the dominance of external summative testing. (para 6.4)
- Where students need to generate evidence during a programme and have it assessed teachers tend conflate formative and summative functions of internal assessment. (para 6.4)
- It has been said that continuous assessment cannot function formatively when each attempt or piece of work submitted is scored and the scores are added together at the end of the course. This practice tends to produce in students the mindset that, if a piece of work does not contribute towards the total, it is not worth doing. (para 6.4)
- We cannot say that any test or instrument sourced externally will automatically be summative and that any piece of less formal assessment done in a classroom will automatically be formative or for learning. Nor can we assume that an instrument developed for external summative purposes and used uncritically or inexpertly will satisfactorily perform a useful formative role. (para 6.5)
- We cannot have good formative assessment in situations where teachers' efforts are not valued, respected and trusted. (para 6.6)
- There is a need to get formative processes right first since it is a prerequisite for both good learning and good summative assessment. (para 6.6)

- A resolution of the tension between formative and summative assessment will require teacher expertise in assessment that is concerned with principle and purpose and not just with process. (para 6.5)
- We should not under-estimate the levels of professional expertise in assessment required of teachers if they are to become key players in the resolution of the tension between formative and summative purposes for assessment rather than just implementing externally controlled initiatives. (para 6.7)

6.22 In relation to the conduct of formative and summative assessment:

- Teachers find it difficult to break out of a mode of assessment that is concerned with making summative judgements. (para 6.8)
- Teachers' understanding of what counts as valid evidence becomes the framework for the knowledge that students have transmitted to them. Students do not tell teachers what they actually know, but what is seen to fit into the frameworks as they understand them. (para 6.8)
- Formative assessment processes may be convergent (such as finding out whether a students knows a predetermined theory) or divergent (seeking to discover what the students knows). What is required is an approach to assessment that enables both convergent and divergent assessment to be pursued at appropriate times. (para 6.10)

6.23 In relation to student motivation:

- There are actions that teachers and schools can take to ensure that the benefits of summative assessment can be had without negative impact on students' motivation for learning. These include
  - promoting and engaging in professional development that emphasises learning goals and learner-centred teaching approaches to counteract the narrowing of the curriculum
  - emphasising learning rather than performance goals with students and developing students' understanding of these
  - providing feedback in relation to these goals
  - developing policies that ensure that the purpose of all assessment is clear to all involved
  - presenting assessment as a process producing results that have to be regarded as tentative and indicative rather than definitive. (para 6.12)
- Students have become very utilitarian in their view of what it is worthwhile to pursue in complying with assessment requirements. This may not exclude enthusiasm for work, and may represent a commitment to the achievement of some long-term goal. (para 6.14)

6.24 In relation to feeding back the outcomes of assessment:

- Even when teachers provide students with valid and reliable judgements about the quality of their work, improvement does not necessarily follow. (para 6.15)
- Feedback should be clearly linked to the learning intention so that the student understands the success criteria and standard. (para 6.16)

- Feedback focuses on the task and not the learner, gives cues about next steps, challenges, requires action and is achievable. (para 6.16)
- Improving the quality and extent of feedback may not be incompatible with the use of internal assessment for summative purposes and may become a feature of more sophisticated management of external assessment through the use of ICT. (para 6.17)
- Feedback in relation to summative assessment that does not include comments and information related to the specifics of the work that students have done is of limited value. Students need to be able to take action that will bring particular pieces of work up to the required standard. For students, feedback that is insufficiently focused is of little use. (para 6.18)
- There are links between quality of feedback and students' capacities to manage their own learning. (para 6.19)

7 The validity and reliability of internal assessment

The reliability of assessment for internal components of examinations

7.1 It has been a popular view that, where qualification outcomes depend on internal assessment they should be accorded a lower status, and where high status is required internal assessment should be limited in its scope and in the weight of its contribution to the whole qualification and should be tightly specified. Moreover, the basis for assessment by teachers should be as explicit as possible and there must be checks that will ensure that differences between assessments made by different people are minimised. In this section we shall review some of the evidence of the reliability of internal assessment in the various contexts in which teacher judgements occur, in an attempt to determine whether there is solid evidence that it is untrustworthy (or at least any more untrustworthy than external assessments).

7.2 In a review of the reliability of examinations Wilmut *et al* (1996) observed that the limited evidence available about the reliability of school examination coursework assessment suggested that the picture was mixed, but not too bleak, and certainly not invariably worse than with some types of written examination. They reported early work by Hewitt (1967) who showed reasonable correlations between assessments made by teachers and an external moderator in O level, and work by Cohen (1974) that returned reasonable inter-assessor coursework reliabilities in a range of CSE subjects. On the other hand, Willmott & Nuttall (1975) did not attempt to determine reliabilities for teacher-assessed components within the O level and CSE examinations that they studied, but doubted that these would be very high. Although the increase in the examination time and scope would normally be expected to result in an increase in reliability, they believed that what they called the 'subjective nature' of internal assessment and the many teachers involved would offset this gain. However, Taylor (1992) reported very creditable correlations in the region of 0.87 to 0.97 between pairs of moderators marking coursework folders in English and mathematics, although the teachers for whom they were responsible may not have been as consistent.

7.3 Commenting on the attitude to reliability in examining boards Wood (1991) has said

> "The whole history of external examining is a tale of curbing and reining in heterodox educational practices, in the interests of fairness to all. Whatever is done, however partial, should be done to all in the same way. Nothing should be done to jeopardise reliability."

Awarding bodies cannot be expected to neglect reliability – they feel obliged to maximise it, as a matter of their accountability to students, teachers and other stakeholders. Wood (1991) argues that, even if awarding bodies' need to focus on reliability places a limit on validity,

> "… it is simply no use having the most wonderful examination in the world if it cannot be marked and graded reliably. The experiences may have been good but what is the point if the rewards are distributed haphazardly?"

So the awarding bodies cannot act as agents for educational reform; this must come from a clearer statement of how we are going to achieve a satisfactory trade-off between validity and reliability in order to achieve more authentic assessment.

Factors affecting the reliability of teacher assessment

7.4 The early work cited above treated teacher assessment as a single outcome and did not include a consideration of the differences in reliability between assessments based on a single piece of work or many pieces of work, made analytically or holistically, made on the basis of compensation or mastery, and so on. These distinctions, indicated in the taxonomy in Section 3, are of crucial importance when deciding how an internally assessed component may be specified and in deciding which approach is likely to lead to an acceptably reliable outcome.

7.5 There is a very substantial literature, covering more than a century of research, on the quality of marking by teachers, both in relation to classroom activities and in relation to examinations and tests. It covers primary, secondary and higher education, and some is clearly appropriate to a discussion of the role of the teacher as assessor in internal assessment within qualifications[4]. In particular we can see that any set of marks can conceal the operation of all sorts of covert rewarding, whether it be gender-related or takes some other form, such as substituting perceived ability as a proxy for achievement (Wood, 1991). Evidence for the effect on assessment decisions of prior knowledge of a student comes also from higher education, and usefully from a review of the social psychology literature, undertaken by Archer & McCarthy (1988) who conclude that

> "… it would seem foolish for psychology teachers to deny the applicability to their own concerns of a huge and varied accumulation of social and psychological evidence, all pointing to the potential distortions and injustices that can result when performances are judged with reference to a body of prior knowledge."

7.6 Wood also reviewed work by Pedulla, Airasian & Madaus (1980) and Kellaghan, Madaus & Airasian (1982) which found that teachers' judgements of students' IQ, English and mathematics performances were confounded with their judgements of other matters such as attention span and persistence, but not with their social behaviours. However, when teachers were given access to test results they more often raised than lowered their judgements of students' capabilities, suggesting that they were susceptible to external influences.

7.7 Similarly, but from assessment within the national curriculum, is evidence of over-marking by teachers, done (perhaps ineffectively and misleadingly) in order to encourage and motivate 12-year-old students (Spear, 1989). Earlier Wood and Napthali (1975) had reviewed the literature then available, and concluded that there was strong evidence of an interplay between teachers' expectations and students' performances, and that the outcomes of assessment would depend on these, but differently from teacher to teacher. The authors were able to generate six derived constructs which were likely to differentiate between students; these were, in no particular order, the

- involvement of the student in the learning situation
- ability that the student has in the subject
- overall ability of the student
- student's behaviour

[4] Wilmut *et al* (1996) included a comprehensive bibliography of sources on marking reliability as a supplement to their main review paper.

- quality and tidiness of the work presented
- interest which the student displays in the subject.

They concluded that the situation may be improved by the use of assessment instructions which relate directly to constructs of this type, though the benefits may be limited.

7.8 There is quite a good deal of international evidence of the effects of gender and other factors on assessment outcomes. In relation to the effects of gender in examinations, there are several studies (such as those done by Murphy, 1978 & 1982) that suggest gender effects based on the type of task set. Carter (1952), amongst others, has shown differences in classroom assessment decisions, depending on the genders of teachers and students. Kelly (1988), on the basis of a meta analysis of 89 independent sets of data drawn from studies in the United States and Europe, suggested strong evidence of a much broader difference in treatment of males and females by teachers. Other work (Husbands, 1976; Branthwaite *et al*, 1981) has focused on the effects of the social interactions and differences in ideological stances between teachers and students, and found these to have some effect on assessment outcomes, and other studies have drawn attention to factors such as the neatness and handwriting in student presentations. In a review of more recent work on gender differences in assessment Stobart *et al* (1992) discussed somewhat conflicting and confusing evidence from a number of smaller studies done shortly after the introduction of GCSE. Here differences in coursework performances between males and females, differences between subjects and differences between students from various schools were identified, but with few clear trends.

7.9 Wood (1991) suggests that the research evidence as to whether teachers can distinguish industry and effort from achievement is unclear. But there is evidence of different outcomes related to a number of non-achievement variables other than gender such as class, physical attractiveness, perceived IQ, and teachers' prior expectations. However, in the area of social class an early small-scale primary school study by Murphy (1974) suggested that there was little evidence for either a halo effect or a self-fulfilling prophecy acting to determine teacher assessments:

> "… explanations of working-class under-achievement by reference to teacher expectations and typifications appear to stand in need of qualification."

7.10 There is, of course, the matter of the interpretation of the assessment requirements, either as they are embodied in a specification or syllabus, or as they are interpreted in the tasks given to students. In vocational qualifications successive refinements in the standards have not necessarily induced greater clarity. Murphy (1995) has discussed evidence concerning the differences in school students' perceptions of tasks, because of the different contexts and circumstances in which they have learned. We are wrong to assume that meaning is inherent in the words used to communicate assessment tasks, and wrong to assume that problems can be given ready-made to students. Assessment outcomes must be interpreted in relation to individual contexts and circumstances. Whilst this conclusion appears to undermine attempts to present the same task to all students, it may support the view that there is a practical limit to the reliability of assessments that can be made through the unmediated delivery of external tasks.

7.11 Thomas *et al* (1998) conducted some work in the context of the authentic testing debate, looking at the extent to which different types of national curriculum assessment serve to support the elimination of bias, or whether the bias is a consequence of the conditions of

learning and it doesn't matter what type of assessment is used. They worked at key stage 1, and used multi-level modelling to look at group differences amongst teacher assessing both standard tasks and classroom work (as part of teacher assessment). The interest was in whether the use of the tasks informs and shapes teacher assessment, and it appears that this is the case to a considerable extent, so that the task results explain the majority of the variance. But there was evidence of differences in interpretation of criteria amongst teachers (varying from subject area to subject area), and of some variance attributed only to teacher assessment. This suggested that either the standard tasks and teacher assessment domains are not identical (as they were intended to be) or that teachers introduce bias into the assessments.

7.12 In relation to competence-based assessment in Australia Maxwell (1997) discussed the basis for teacher judgements in schools that ranged across the States from binary judgements of competence/non-competence through 3, 5, 8 and 10 level systems. In Queensland at senior secondary level there are centrally developed syllabuses implemented by schools which develop accredited programmes of work, leading to progressive assessment over 2 years, using a variety of tasks. Assessment criteria are typically very detailed. Moderation procedures are managed internally by the schools and then managed by the board between schools, where expert panels sample portfolios of work. An earlier study had suggested high levels of agreement at the inter-school level. Maxwell's study focused on teacher decision-making, exploring this in depth for a small number of teachers in English. Typically

- a teacher started with the most holistic judgements on an item of work for a student, before referring to the criteria, then shuttled to and fro between the holistic judgement and the criteria, in a process of refining the decision
- a teacher referred to other students' work in a further process of refinement
- there was often a need to trade-off among criteria, because of inconsistencies in performance (note here the similarity to the results from the work on grade criteria in the UK; see Cresswell, 1987a); this forced teachers to consider inter-relationships between criteria, how they may be weighted, and how they may be aggregated
- where standards were not sufficiently explicit a teacher distilled these from common characteristics across a range of students
- the links between particular criteria and the overall standard for a level were often poorly articulated
- conflicts arose when criteria designed to apply to all syllabus implementations were applied in the context of a specific implementation; the syllabus was used to reconcile these, particularly because this is what would be used as a basis for decision in inter-school moderation.

7.13 We must conclude that teacher judgements are complex and that their interpretation of evidence goes beyond simply matching performance to description (Hager & Gonczi, 1993; Wolf, 1995). The process requires social construction of standards involving deliberation and interpretation (Radnor & Shaw, 1995) and the interpretation of evidence involves implicit models of assessment practice (McCallum *et al*, 1995). It is generally thought that the equivalent application of standards (across, tasks, assessors, students, institutions and time) depends on shared experience and understanding

amongst those involved, and depends on teachers having a clear perception of the decisions they are making.

Teachers making predictions of examination results

7.14 There has been some interest in the extent to which teachers can make accurate predictions of examination results (a task that some awarding bodies have sometimes required them to do). It has been said that any failure on their part to do this is evidence of a lack of expertise in assessment, but this is clearly a rather simplistic view. Petch (1964) did one of the earliest and most celebrated studies of teacher estimates of examination results and how these related to the actual grades awarded. Generally the best predictions were for languages and mathematics, with the lower ones for humanities and English. There were generally bigger differences between schools within a subject than between subjects and, over all subjects and schools, there was grade agreement in about 43% of cases, but the examination grade was higher than the teacher estimate in 18% of cases, but lower in 39% of cases, sometimes heavily so. He also looked at the methods used by schools to arrive at the estimates. These turned out to be very diverse so that, for example, some schools used mock examinations, but others not. Petch pointed out that, where estimates and results differ, both may be 'right' since the conditions are different.

7.15 Murphy (1979) did another of the early studies, and looked at teachers' predictions of O and A level grades, finding a reasonably high level of agreement (80% accurate or within one grade at O level and 67% at A level). This rather confirmed earlier work by Petch but, when Murphy looked at rank orders on a centre by centre basis, some correlations between teacher rank orders and examination mark rank orders were very low, though with an overall average value of 0.66. He concluded that teachers and external examinations were not operating on an exactly similar basis; Wood (1991) later suggested that Murphy is saying that the more teacher assessment is taken out of the control of external examinations, the more they will use their own criteria for making judgements, and that the greater the differences will be between these and examination results.

7.16 Much more recently Delap (1995) used a multi-level model for an analysis of eleven A level syllabuses for which teachers in a sample of schools were asked for estimates, specially for the study. Overall results were similar to previous studies: 30% of teachers got the grade right and 72% were correct or within one grade, though 25% estimated a grade high and 17% estimated a grade low. Differences between subjects and through the grade range were observed, but were difficult to explain.

Reliability of teacher assessment in vocational qualifications

7.17 There is much more limited evidence about the reliability of assessments made in vocational qualifications, many of which are characterised in the UK by being related to concepts of competence. As part of the review of NVQs conducted for the Beaumont report a small scale study was conducted by a team at the University of Nottingham (1995). It included two parts: a study of the degree of agreement between assessors in independently viewing portions of portfolios in 5 NVQs in different sectors, and a series of observations of assessments undertaken in colleges, training centres and workplaces. At the level of holistic judgements of the evidence submitted assessors generally agreed on whether candidates were competent but disagreed over details, particularly over whether the evidence presented was sufficient. In some cases the evidence suggested

competence, but was insufficient to give the assessor confidence in this judgement. However, there were evident dangers that the candidates' assessors would be easily swayed by the quantity and presentation of evidence rather than by its substance, and doubts were expressed about the authenticity of some of it. Many of the candidates' assessors appeared to exercise a strong control over the evidence to be placed in the portfolio, and there were many standard exercises.

7.18 In this work assessors were clearly making global rather than specific judgements, and Helsby *et al* (1998), who evaluated Advanced GNVQ practice in 12 schools and colleges, discussed the nature of the assessment process in interviews with students and staff. The important point (reinforced by many other writers) was that the highly specified, complex and atomised structure of the specifications creates a fractured learning experience, with competence decided on a piecemeal basis, rather than by taking a holistic view. Wolf (1995) has taken up many of the issues surrounding assessment in these settings, and commented that

> "The inherent variability of the contexts in which competence is tested and displayed means that assessors have to make constant, major decisions. They must determine how to take account of context when judging whether an observed piece of evidence fits a defined criterion. In other words, they operate with a compensating procedure which itself requires and internalised holistic model – not a simple set of atomised domain descriptors."

7.19 She also discussed some further work on the reliability of assessment in vocational qualifications (Wolf, 1998), summarising findings by Eraut *et al*, 1996; Pedreschi *et al*, 1994; Raggatt & Hevey, 1995; Wolf *et al*, 1994; Broadfoot *et al*, 1995; all of which suggested that reliability is low and that inconsistent judgements were commonplace, a view also taken in a number of contemporary inspection reports. A later review of GNVQ produced rather better results, although Wolf criticised the interpretation of these as over-optimistic.

7.20 However, it is a moot point whether the disagreements between assessors in these studies are larger or smaller than we would get in some essay marking or in some comparable coursework assessment. We won't know until some further research is done, but modest reliabilities have been a feature of many assessment processes, unless they are extremely tightly controlled and relatively narrowly focused. It has been said that the present system of verification is an inadequate quality control for reliable assessment and that systems of moderation provide more direct controls at the output stage.

7.21 Of course, workplace assessment is only valid when it registers accurately the presence of skills which convert directly into occupational competence (Wood *et al*,1989). Studies of the reliability of the assessment are of little consequence if the validity is so limited that the competencies acquired in the course of an NVQ are not transferable into the relevant workplace setting (Tolley *et al*, 2003).

Interactions with issues of validity

7.22 It is difficult to discuss the reliability of internal assessment in isolation, and without some overview of its validity, simply because so many decisions about its implementation depend upon perceptions of its dependability (usually in comparison with written tests and examinations). However, many of the validity issues are covered

elsewhere in this review, so what is included here is a discussion that introduces some additional sources of information that relate validity to reliability.

7.23 At a quite early stage in the development of internal assessment in school examinations the conventional view that internal assessment should only be introduced if it could be shown that written examinations were not up to the job was challenged. Nuttall (1981) took the view that external written examinations should be seen as a sub-set of what may be assessed in school-based assessment, and linked this to the view that it was misleading to report a teacher-assessed component separately. To do so is an artifice of the assessment method and not a representation of what is assessed.

7.24 In a later and much wider discussion he viewed validity as the fidelity of the inference that can be made (as a result of an assessment) to the universe of the behaviour that is of interest (Nuttall, 1987). In workplace settings, for example, where the assessment (which samples workplace tasks) is conducted in the same context as the workplace performance, validity is high. Here, the use of performance assessments and simulations will provide a much better prediction than, say, paper and pencil tests. He then discusses how performance may provide valid predictions of competence, and the circumstances under which this validity may increase. Clearly these include a variety of factors arising from the relationship between the assessor and the learner, and the ways in which the assessment is conducted. They also include the context in which the assessment is conducted, and the improvement in performance that may result when the context is appropriate. He commented:

> "It is often concluded that teachers are overgenerous in their assessments; an alternative conclusion is that the conditions under which course work was conducted and assessed elicited or facilitated genuinely better performance."

The emphasis on context is of central importance:

> "Assessment (like learning) is highly context-specific and one generalizes at one's peril."

He argued that assessment tasks and situations lead to best performance when they are concrete and within the experience of the student, clearly presented, seen to be relevant to the learner's current concerns and use situations which are not unduly threatening and where there is a good relationship with the assessor.

7.25 In his discussion of the same theoretical issues Wiliam (1993) described domains of assessment and defines

- validity as the extent to which inferences within and outside the domain of assessment are warranted; this subsumes
- dependability as the extent to which inferences within the domain of assessment are warranted, and which subsumes
- reliability as the extent to which inferences about the parts of the domain actually assessed are warranted.

7.26 There are considerable difficulties in defining the domains, even when apparently precise criteria are in use. In fact, these can only be operationalised through the use of norms, which define the boundaries between mastery and non-mastery. Wiliam discussed whether a higher level sub-domain of an attainment target subsumes a lower one, and concluded that this is not automatically the case: some pupils can achieve, say,

level 3 on a level 3 test but not level 2 on the level 2 test. He noted that this isn't the case with teacher assessment, where teachers will not deem a pupil to be at level 3 without being sure that they have achieved level 2, adapting their interpretations of the domains in order to ensure this.

7.27 Returning to issues of context, there is no point in making the assessment contextualised if the instruction is not contextualised, either by the teacher or by the student, says Cumming (1997). It is instruction which leads the process, and

> "… authenticity could be defined as assessment which is appropriate to the purpose of the instruction and outcomes to be observed .. learning and instruction are complex and authentic assessment as pseudo-real world performance may not always be the most valid."

7.28 Some further insights come from studies of the behaviour of teachers as assessors in specific situations. An example is the work of Morgan (1996), who described detailed and structured work with teachers, who were asked to make assessments of mathematics coursework materials from students whom they didn't know, and then discuss these in detail. She identified a range of strategies they used for assimilating and judging the coursework; sometimes an individual teacher will change strategies during the assessment of an individual piece of work. This arises from the tensions in the roles of teachers, who are simultaneously agents of the awarding body in making assessments, concerned to further the learning of the individuals whom they are assessing, and anxious to show the work of these students in the best possible light. Teachers often sought to build up a picture of the student and his or her understanding, as an essential step in making an assessment judgement. However, this may be a judgement that is of the student's perceived ability, which may result in the teacher overlooking omissions or inferring performances, in order to arrive at what is perceived to be a fair judgement. In the case of this study, the inferences of ability were made entirely from the text of the coursework, using criteria that were derived from their own experience. This led in some cases to them assessing by making comparisons between what they were looking at and what they regarded as ideal pieces of coursework. However, where the pattern of the coursework was unconventional the different strategies produced different judgements. Further training in assessment, if it were provided, would tend to reinforce conventional responses, and it was possible that the purposes and scope of the coursework could only be clarified through a process of agreement trialling.

7.29 The extent to which our concepts of validity have lagged behind assessment developments is the subject of an extensive and complex paper by Moss (1992) who explores the need for a reconceptualisation of validity in relation to the increasing use of performance assessment (much of which is conducted by teachers). Writing of education in the United States, she notes that most validity enquiry has been based on scientific models linked to behaviourist principles, and these have tended to emphasise standardised and privileged forms of assessment

> "... in order to enhance reliability, generalisability and comparability of scores".

The growth of authentic assessment has led to the valuing of less standardised performance assessments, where teachers and students may develop their own tasks, develop criteria for assessing them and discuss the interpretation of assessment outcomes with stakeholders. This embraces some portfolio assessment in the United States and comes close to some aspects of portfolio assessment in Britain. She considers

that the gains from such assessment are such that we need to expand concepts of validity so as to consider the social consequences of the interpretation and use of information from less standardised performance assessments; at the same time we need to be more aware of the effects that the use of privileged forms of assessment have on society.

7.30 In a review of the evidence of the reliability and validity of assessments used by teachers in education from 14 – 18 for summative purposes Harlen develops conclusions on the basis of an analysis of 30 research sources selected through an EPPI review process (Harlen, 2004). The selected sources were chosen because of their relevance to the area of study, soundness of research methodology and the weight that could be given to the evidence produced. Findings were presented in relation to a main question concerning evidence for the validity and reliability of internal assessment and a subsidiary question concerned with the factors affecting validity and reliability. Most of the findings do relate to a specific student age group, context or subject, and Harlen rightly warns about generalising from specific findings to the whole area of internal assessment.

7.31 Some of the sources used have been discussed elsewhere in this review, although the EPPI study has tended to focus on assessment within the national curriculum, or equivalent overseas provision, rather than qualifications in the 14-19 sector. Additionally it does not include non-research evidence or sources where the evidence is less strong. The principal findings relevant to this review and relating to the main question are as follows.

- The reliability and construct validity of portfolio assessment where tasks were not closely specified were low (but see the later discussion on portfolio assessment).
- A finer specification of criteria, describing progressive levels of competency, is capable of supporting reliable teacher assessment whilst allowing evidence to be used from the full range of classroom work (Rowe and Hill, 1996).
- There is considerable error and bias in teacher assessment of different groups of primary students but the interpretation of correlations of teacher assessment and standard task results should take into account variability in the administration of the standard tasks.
- 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, though there was a later decline in earlier collaboration among teachers and sharing interpretations of criteria.
- Results of some teacher assessments and standard tasks agree to an extent consistent with the recognition that they assess similar but not identical achievements.
- The clearer teachers are about the goals of students' work, the more consistently they apply assessment criteria and 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.
- When rating students' oral proficiency in a foreign language, teachers are consistently more lenient than moderators, but are able to place students in the same rank order as experienced examiners.

- Teachers are able to score hands-on science investigations and projects with high reliability using detailed scoring criteria and teachers' assessment of practical skills in A level science makes a valid contribution to assessment but with little evidence of generalisability of skills across subjects.
- Teachers' perceptions of students' ability and probability of success on a test are moderately valid predictors of performance on the test, as are student self-assessments of their performance on a test after they have taken it.

7.32 On the same basis, findings related to the subsidiary question were

- There is bias in teacher assessment relating to student characteristics, including behaviour (for young children), gender, special educational needs, overall academic achievement and verbal ability; this may influence judgement when assessing specific skills.
- There are differences among schools and teachers in approaches to conducting teacher assessment within the national curriculum.
- Differences between subjects in how teacher assessment compares with standard tasks or examinations results have been found, but there is no consistent pattern.
- It is important for teachers to follow agreed procedures if teacher assessment is to be sufficiently dependable to serve summative purposes.
- Training for teachers to improve the reliability of their assessment should involve them in the process of identifying criteria so as to develop ownership of these and an understanding of the language used. Training should also focus on the sources of potential bias that have been revealed by research.
- Teachers can predict with some accuracy their students' success on specific test items and on examinations. There is less accuracy in predicting A level grades.
- Detailed criteria describing levels of progress in various aspects of achievement enable teachers to assess students reliably on the basis of regular classroom work
- Moderation through professional collaboration is of benefit to teaching and learning as well as to assessment. Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others can give.

Summary

7.33 In relation to what we know about the reliability of teacher assessment:

- The evidence is typically very mixed. (para 7.2)
- Differences between subjects in how teacher assessment compares with standard tasks or examinations results have been found, but there is no consistent pattern. (para 7.32)
- A popular view has been that where qualification outcomes depend on internal assessment they should be accorded a lower status and where high status is required, internal assessment should be limited in its scope and weight and should be tightly specified. (para 7.1)
- Awarding bodies cannot be expected to neglect reliability. (para 7.3)

- Much of the research evidence does not include a consideration of the differences in reliability between assessments based on a single piece of work or many pieces of work, made analytically or holistically, made on the basis of compensation or mastery, and so on. These distinctions are of crucial importance when deciding how an internally assessed component may be specified and which approach is likely to lead to an acceptably reliable outcome. (para 7.4)
- Any set of marks can conceal the operation of all sorts of covert or unintended rewarding, including that which is gender-related, substituting perceived ability as a proxy for achievement, rewarding diligence or tidiness or reflecting expectations of performance. (paras 7.5 – 7.9)
- There is considerable error and bias in teacher assessment of different groups of primary students but the interpretation of correlations of teacher assessment and standard task results should take into account the variability in the administration of the standard tasks. (para 7.32)
- The reliability and construct validity of portfolio assessment where tasks are not closely specified is low. (para 7.32)
- It is wrong to assume that meaning is inherent in the words used to communicate assessment tasks, and that problems can be given ready-made to students. Assessment outcomes must be interpreted in relation to individual contexts and circumstances. Consequently there is a practical limit to the reliability of assessments made through the unmediated delivery of external tasks. (para 7.10)
- A finer specification of criteria, describing progressive levels of competency, is capable of supporting reliable teacher assessment whilst allowing evidence to be used from the full range of classroom work. (para 7.32)
- In general, vocational qualifications studies have suggested that reliability is low and that inconsistent judgements are commonplace. (para 7.19)
- It is not clear whether the disagreements between assessors in studies of vocational or occupational qualifications are larger or smaller than in some essay marking or in some comparable coursework assessment. (para 7.20)

7.34 In relation to the methods that teachers use to arrive at assessment:

- Teacher judgements are complex and their interpretation of evidence goes beyond simply matching performance to description. The process requires social construction of standards and the interpretation of evidence involves implicit models of assessment practice. (para 7.13)
- It may be the case that a teacher: starts with the most holistic judgements before referring to the criteria, then shuttles to and fro between the holistic judgement and the criteria in a process of refining the decision; may refer to other students' work in a further process of refinement; may make a trade-off among criteria, because of inconsistencies in performance; may distil standards from common characteristics across a range of students; resolves conflicts that arise when criteria designed to apply to all syllabus implementations are applied in the context of a specific implementation. (para 7.12)

- The equivalent application of standards depends on shared experience and understanding amongst those involved. (para 7.13)
- Teachers who have participated in developing criteria are able to use them reliably in rating students' work. (para 7.32)
- Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others can give. (para 7.32)
- The more teacher assessment is taken out of the control of external examinations, the more they will use their own criteria for making judgements, and the greater the differences will be between these and examination results. (para 7.15)
- In vocational and occupational qualifications assessors may agree on whether candidates are competent but disagree over specific judgements. There are dangers that assessors are easily swayed by the quantity and presentation of evidence rather than by its substance. (para 7.17)
- The variability of the contexts in which competence is tested and displayed means that assessors have to take account of context when judging whether an observed piece of evidence fits a defined criterion, operating with a compensating procedure which itself requires an internalised holistic model and not a simple set of atomised domain descriptors. (para 7.18)

7.35 In relation to issues linking reliability with validity:

- External written examinations should be seen as a sub-set of what may be assessed in school-based assessment. (para 7.22)
- Where the assessment is conducted in the same context as the performance, validity is high. Assessment is highly context-specific and one generalizes at one's peril. Assessment tasks and situations lead to best performance when they are concrete and within the experience of the student, clearly presented, seen to be relevant to the learner's current concerns and use situations which are not unduly threatening and where there is a good relationship with the assessor. (para 7.24)
- Teachers will not deem a pupil to be at level 3 without being sure that they have achieved level 2, adapting their interpretations of the domains in order to ensure this. (para 7.26)
- Teachers often sought to build up a picture of the student and his or her understanding, as an essential step in making an assessment judgement. However, this may be a judgement that is of the student's perceived ability, which may result in the teacher overlooking omissions or inferring performances, in order to arrive at what is perceived to be a fair judgement. (para 7.28)
- The growth of authentic assessment has led to teachers and students developing their own tasks and the criteria for assessing them and discussing the interpretation of assessment. Authentic assessment requires that we reconceptualise concepts of validity. (para 7.29)
- Workplace assessment is only valid when it registers accurately the presence of skills which convert directly into occupational competence. Studies of the reliability of assessment are of little consequence if validity is so limited that the competencies acquired are not transferable into the relevant workplace setting. (para 7.21)

8 Quality assurance and control of internal assessment

Some broad descriptions

8.1 In the context of this discussion of internal assessment, quality assurance may be seen as the processes carried out in order to ensure that assessments meet the requirements of validity and reliability, normally carried out internally in the form of internal standardisation, moderation or verification. Quality control may be seen as the external checking process, carried out across centres by an awarding body, in the form of external moderation or external verification. Put another way, control over internal assessment can be provided before it starts, as part of its process, or by looking at its outcomes; just about every combination of these approaches has been used.

8.2 Much of what is relevant here is concerned with the processes used to manage internal assessments made at a local level. As a starting point we can divide the approaches into moderation (where the outputs generated by learners are inspected and/or adjusted – Elley & Livingstone (1972) called it "any method of determining the differences in attainment of various groups of pupils" – and verification where the systems used for the management of the internal assessment are checked). There can be internal and external components to both processes. We should not get too prescriptive about these terms: quality assurance and control mechanisms often involve a mix of both methods, but it is important to see that they do represent approaches that lead in different directions and offer different possibilities – the ideas are certainly not interchangeable.

8.3 It seems always to be the case that teachers (and, through them, students) interact very closely with whatever methods of quality assurance and control are implemented in school, college or workplace. Ecclestone & Hall (1999) comment that

> "… teachers rely on a complex blend of broad experience, good relations with colleagues and awarding body officers, and informal sharing of assessment judgements to arrive at a notion of what constitutes a 'standard'"

and that consequently

> "… students show high levels of trust in their teachers' ability to assess competently and fairly".

8.4 In relation to further education colleges they commented that FEFC inspectors saw quality assurance and quality control as a holistic set of practices within supportive institutions which maintain a balance between flexibility, creativity and meeting individual students' needs and requirements to conform to national standards. They drew a distinction between academic awarding bodies that saw their role as supporting and guiding teachers to undertake quality assurance and control whilst vocational bodies had stronger systems for intervention and regulation of what teachers and assessors do.

8.5 What also has emerged in recent years is an approach to the quality assurance of assessment that regards this as one aspect of the management of all aspects of quality across an institution. This is particularly the case in further education colleges and in relation to vocational qualifications where, with the support of some awarding bodies, approaches to the management of quality attempt to address, in a coherent way, issues concerned with the quality of learning and teaching, staff support and development, resource allocation, student recruitment and retention, management effectiveness and

the conduct of assessment. Thus, processes such as internal verification become networked within wider quality considerations. The characteristics of such systems are that they

- are centrally managed at a senior level
- incorporate development or business plans and internal monitoring and feedback processes that feed into planned response mechanisms
- are designed to cut across structures that separate departments or faculties
- require accurate methods for information and data gathering and dissemination
- operate on a continuous basis (Wilmut & Macintosh, 2001; Wilmut & Murphy, 2001).

8.6 At the same time Wood (1991) (citing earlier work by Petch, 1963) considered that there is often over-moderation, and that a weight for internal assessment within an examination of less than 25% will not make much difference to the final grades except in the cases of a very small number of individuals; moderation involving scrutiny is too expensive in these cases. The solution would be to use spot-check moderation, or to monitor using statistical methods and scrutinise only occasionally or when danger signs appeared. This and the potential for enhancing internal quality assurance methods does lead into later discussions of the operation of verification and moderation, the roles of verifiers and moderators and the possibility of accrediting centres to conduct internal assessment with limited levels of supervision, either because of satisfactory past performance or because of the existence of appropriate quality assurance and control mechanisms.

Providing prior information and support for teachers and assessors

8.7 Since moderation and verification procedures can be rather expensive, differences in approaches by teachers and assessors might be prevented if they were given the right sort of information or support beforehand. It might even be possible to accredit them to conduct assessments, within an established framework, and subject them to periodic checks. We can identify three general support strategies that might be adopted.

- Teachers are provided with materials designed to support their assessment, often in the form of exemplification.
- Training is provided in the operation of particular assessments.
- Steps are taken to develop teachers' understanding of the principles of assessment.

8.8 Most of the support provided in the past has been a mix of the first two of these, with training programmes aimed at specific aspects of assessment (a good example is the Key Skills Support Programme) or they have been provided with materials such as the optional tests and exemplification materials available for the national curriculum. There has recently been a considerable increase in the amount of material available on the internet and evaluations suggest that this has been widely used and generally been well regarded (Harlen *et al*,1993; Emery *et al*, 1998a; Emery *et al*, 1998b; Gipps & Clarke, 1998a). It appears to be strongly influencing the approaches to assessment throughout the key stages. It is not uncommon for support also to be provided through the use of banks of assessment materials that teachers can download. Walker (1979) proposed this in Scotland, Masters (1986) described it in Australia, Black (2004) has emphasised its

value in school science, and the Department for Education and Skills (DfES) is currently investing resources in a bank of materials for the assessment of key and basic skills in the UK. The risk with these materials may be that their formative and summative purposes are casually conflated by their designers and/or users.

8.9 Part of the process of moderation-free teacher assessment in the national curriculum in the mid-1990s depended on encouragement by the School Curriculum and Assessment Authority (SCAA) and then QCA for school departments to develop portfolios of exemplar materials to be used as a reference against which to check subsequent assessment decisions. Evaluations suggest that these had become quite widely used but that their compilation was time consuming (Taylor & Lee, 1994) and it is said that their use has reduced somewhat as teachers devote less time to internal moderation meetings and there are fewer opportunities for inter-school discussions.

8.10 There are also, of course, many possibilities for training or supporting teachers in making assessments and awarding bodies and others have undertaken considerable amounts of work of this kind. It is possible that too much of this training has been in the provision of short programmes, that are too focused on specific qualifications and that do not provide the continuity or depth that is required for real professional development. However, to do this would be costly and from time to time schools, colleges and training providers have developed networks to support teachers' professional development in assessment. These are strongly in tune with calls for a shift away from post-assessment adjustments, an increase of emphasis on improving assessor skills, signalling standards and moderating comparability of assessment activities before they are used, and assessor accreditation and a greater emphasis on professional judgement (Strachan, 1997) but are expensive and time-consuming to operate.

Approaches to moderation

8.11 There have been various ways of describing approaches to moderation. Cohen & Deale (1977) and Walker (1979) identified three types of moderation: by inspection, statistical and consortium. They described what they saw as the desirable features of schemes of teacher assessment and moderation, namely that

- teachers should be given training, with opportunities for trial marking and discussion
- moderation feedback should be provided
- moderators should themselves have standardisation and monitoring
- moderation should be done without knowledge of teachers' marks or grades
- double moderation is desirable
- sample sizes for moderation should be adequate to ensure comprehensive judgements
- statistical moderation against written papers as criterion should not be used as a sole method, but should trigger inspection.

Moreover, moderation procedures should be adapted to meet subject needs and assessment procedures should not be constrained to fit moderation requirements.

8.12 A review from the Joint Matriculation Board in the 1970s (Smith, 1978) reported a poll of teachers in a few subjects that had internal assessments and found a big majority in favour of moderation by inspection (which, in the Joint Matriculation Board's (JMB's) case, was the expert re-marking of a sample of material). Then, in the wake of the

publication of the Waddell report, Ward (1982) discussed the difference between continuous assessment (where the assessments were done during the course, over an extended period) or coursework (where they were done at the end) – this is a terminology earlier suggested by Hoste and Bloomfield (1975) – and suggested that the final teacher judgements or summative decisions may be deferred to the end of the course when moderation will take place. It was recognised that there are issues of authentication of students' work and difficulties of moderation when aspects of process or performance are being assessed, in contrast to the moderation of permanent outcomes.

8.13 Buchan's work in 1993 included a discussion of moderation in GCSE science, producing some typical conclusions, including that

- statistical moderation was poorly understood, with scathing criticisms from teachers of the inappropriateness of using written papers as a moderating instrument for work of a practical nature
- whilst the use of back-up moderator visits to schools offset some of these problems, postal moderation remained the most widely used method
- schools often received very little feedback on the reasons for changes to marks as a result of moderation, and there is therefore little basis upon which teachers could modify their practice.

Where teachers acted as moderators for boards their experience was considerably widened, and this could be passed on to colleagues but, in general,

- teachers had operated a system in which the majority were not involved in developing experience of moderation
- moderation procedures were generally not elaborated to teachers
- decisions made within moderation were seen to take little account of the conditions under which teachers worked and how these affected assessments which they make.

8.14 Broadfoot (1994) provided a review of quality assurance and control systems in a number of countries against a background which contrasts the emerging interest in authentic assessment (stressing valid and useful forms of learning, with the need for learning of a broader range of higher-order intellectual skills) with the earlier dependence on reliability as the sole criterion of quality. She judged quality assurance and quality control in relation to utility (how far the assessment procedures contribute to or detract from national educational goals) as well as in relation to their contribution to validity and reliability. She commented on systems of moderation in Australia that, at that time, depended on a combination of devolution of responsibility for assessment design and conduct to teachers, who are given appropriate training, plus moderation provided for by common syllabuses and a combination of in-school, district-wide and system-wide meetings to ensure common standards. Statistical moderation was seen to be arbitrary and not professionally supportive (although it is now more widely used). She commented that

> "Australian school systems have much to teach others when it comes to protecting the validity of the system and the reliability of assessments made within it, at the same time as protecting the rights of individuals."

8.15 The need to exercise tight control over internal assessment is particularly strong in very high stakes situations where an awarding body may feel very anxious about any process that was seen to carry a risk of public disapproval. In such situations a tightly controlled and dependable moderation system is very important and change may have to come from a re-appraisal of the quality of learning and nature of the curriculum rather than from the management of the examination system (Hong Kong Examinations Authority, 1998).

Statistical moderation

8.16 Wood (1972) called statistical moderation an 'armchair' method; it was well established when he wrote and it has remained a mainstay of examination systems until this day, though mostly now used in the UK to signal anomalies that require investigation rather than as a basis for adjustment of teacher assessment marks. It is relatively cheap and easy to use. Wood took the view that it didn't matter a lot that the teacher assessment and the written papers didn't share the same objectives; there is inevitably some overlap that can be used as a basis for moderation, however imperfect. Ward (1982) pointed out that the correlational relationship between the marks on the moderating instrument and those on the teacher assessment often vary widely from centre to centre and is vulnerable to small sample effects. He then discussed the logic of using the method:

> "Any form of statistical moderation using a reference component consisting of a written paper is open to the criticism that the procedure is statistically illogical. If the teacher-assessed component and the reference component have a low correlation because they are assessing quite different examination objectives, then clearly the adjustment of one by relating it to the other is not justified. On the other hand, under conditions where it *is* justified, i.e. the two components are highly correlated, it is questionable whether assessment of coursework is necessary at all. So, on measurement grounds, it could be dispensed with. But, the fact that two components are highly correlated, does not mean that they are necessarily testing the same objectives. Hence there may be good reasons for retaining the teacher assessed component for its contribution to the content validity of the examination."

8.17 Statistical moderation generally involves the scaling of marks whilst preserving the rank order of candidates within a group. Awarding bodies generally require there to be internal moderation that enables them to apply a statistical method to a whole centre rather than to class groups. Where there is evidence that the rank order cannot be relied upon, samples of work may be re-assessed by moderators. Wood (1991) suggests that the assumption of preserving teachers' rank order of marks may not be consistent with the available evidence. He suggests that it is the overall level that teachers find it hard to settle upon, so that they may be biased upwards (see Spear, 1989) and this is often held to be the case.

8.18 Wood also discusses a range of technical issues connected with statistical moderation including those on the effects of small samples and skewed distributions of marks, component weighting (Adams & Wilmut, 1982; Cresswell, 1987b) and suitable statistical models (Good, 1988). Statistics may be used to generate alerts rather than simply as a basis for blanket adjustments and Murphy (1981) favoured this approach, as did Cohen & Deale (1977) and Wilmut (1977). QCA's work on risk assessment (QCA, 1998) headed in the same direction in relation to vocational qualifications, though with a potentially more complex model than was ever used with public examinations, and

alerting is a system which has been used in Victoria, Australia since 1994, for Year 12 examinations, leading to university entrance.

8.19 This last example illustrates a situation in Victoria where the earlier verifications system involved standardisation and expert review procedures, but was judged to be insufficiently reliable and increasingly costly (Brown and Ball, 1992). Whilst simple statistical moderation was seen to be cheap and objective it is weak in situations where objectives covered in the school assessment are not covered in the test and where entries are small. It is also open to manipulation in schools where, in order to maximise students' positions on a tertiary entrance score, teachers concentrate on maximising written paper results, so as to drag up scores on school-based assessment. Hill *et al* (1997) describe the new Victoria procedure that uses a mixture of teacher-marked common assessment tasks undertaken over a period of time alongside school-based tasks assessed according to centrally-developed criteria and supported by school-based systems of support and quality assurance. There are also written examinations to which are now added a general achievement reference test that is administered before the finalisation of school-based assessments and the examination; this follows development work undertaken by Hill *et al* (1993). However, no adjustments to teacher assessments follow: the only outcome is that schools outside a certain tolerance limit are subject to expert review. The method is considerably more complex than has generally been the case with statistical moderation, and would certainly not be understood by almost all teachers, students and most other people, although the authors don't report complaints on this score. They suggest that the new procedures reduce costs and workloads, concentrating resources in schools where moderation is needed and is generally effective.

8.20 Nuttall & Armitage (1983) also worked on creating a moderating instrument, designed to monitor grading standards on Training and Enterprise Council (TEC) awards. The research included developing two types of monitoring test: a broad test, capable of providing assessment relevant to a range of similar units, and specific tests, each relevant to one unit. The former was seen to be easier to manage and cheaper to use, but to potentially carry less credibility with users and practitioners in colleges. The tests were not used alone: it was found that the strongest moderating instruments consisted of the test score, a summary measure of student performance on earlier units, and the student's age. The previous performance was a stronger component of the instrument than the test score, whether broad or specific. The instrument was designed to be used to alert the awarding body to some apparent discrepancies in standards, and not as a basis for adjusting scores.

8.21 In trials, 89% of over 250 classes of students produced results within acceptable limits; of the rest, about half needed some attention, and the rest were felt to be false alarms. Separate scrutiny of a sub-sample of over 50 classes revealed no cases where the instrument failed to produce an alert.

Moderation by inspection

8.22 Until we come to the era of national curriculum assessment there is relatively little literature that reports on the effectiveness of moderation by inspection, although the similarity between the processes and the marking of examination scripts suggests that some of our understanding of the reliability of marking might be applied here. On the other hand there is little direct evidence of the behaviour of moderators or of the processes that they use. Radnor & Shaw (1995) confirm that the publications relating to

the moderation of GCSE coursework have generally not drawn on detailed fieldwork with teachers and schools.

8.23 Early work by Hewitt (1967) and Christopher *et al* (1970) found very reasonable correlations between the grades awarded by teachers and those awarded by the moderator for English Language coursework who re-marked the materials. In contrast James & Conner (1993) sought to get closer to the processes of assessment and moderation, as used in key stage 1. They studied the work of LEA moderators in four LEAs in 1992; these were working under the framework devised by the Secondary Examinations and Assessment Council (SEAC) for this purpose. The authors make a number of observations which illustrate the problems which arise when too wide a range of expectations are attached to a quality control process, which is insufficiently resourced.

- First, the moderators were expected to focus too much on ensuring reliable assessment and not enough on validity. They had insufficient time to enable them properly to investigate validity in teacher assessment and, of course, reliability is easier to control.
- At the same time teachers did not have enough time to discuss issues of the validity of their assessments.
- Moderators were unable, in the time available, to provide the required quality control as well as support for the teachers. Moderators differed in the emphasis which they put on these roles.

8.24 Putting responsibility on schools to ensure assessment quality, monitored by OFSTED (as was subsequently done), is thought unlikely to solve the problem; schools do not have the resources to do this job properly, and OFSTED is unlikely to provide either the depth of scrutiny or the quality of feedback needed in order to improve validity and reliability. Now that moderation in key stage 1 is being revisited some of these issues will probably come up again.

Consensus moderation and agreement trialling

8.25 Consensus moderation was widely used in CSE examinations and has been regarded by some as a component of a golden era of internal assessment, mainly on the grounds that it promoted professional development as well as achieving the required levels of standardisation. It also formed an essential component of the radical reform of assessment in Queensland following the Radford report of 1972, and continues to the present day. That said, Walker (1979) commented that the Dunning report in Scotland reported that Queensland teachers found the moderation burdensome and Butler (1995) reported that, after more than 20 years, science teachers were using their own rather weak versions of external assessments for internal purposes, with little innovation.

8.26 In fact, in the UK, consensus moderation was often done by the CSE boards on a goodwill basis, with many costs actually being borne by schools. Wood (1991) observed that the CSE boards used consensus moderation as a method of operationalising teacher control, which was central to their purposes, and that they frequently combined consensus and expert judgement in local groups.

8.27 There are various terms: consensus moderation, consortium moderation, agreement trialling and group moderation share a common philosophy but not necessarily exactly the same form or purpose. Although the term ‘agreement trialling’ was in widespread

use in the early days of CSE, a distinction has now been drawn between procedures which focus on professional development in assessment and those which focus on making or ratifying decisions about assessment for a specific purpose: the former can be regarded as agreement trialling and the latter as moderation (Wilmut, 1997). Some CSE boards set up group or consortium moderation panels, composed of teachers, whose role sometimes extended to control the written examinations as well as providing moderation for teacher-assessed components. Elsewhere, consensus moderation meetings have been managed by a board-appointed moderator, and attended by all teachers conducting teacher assessment within a particular examination.

8.28 Devotees of this approach to moderation point to its power in relation to both the conduct of the qualification and the curriculum and pedagogy. Thus, Maxwell (1994) also reviews the consensus moderation used in Queensland, and draws on research (by Masters & McBryde, 1993; Allen & Travers, 1995; and Travers & Allen, 1994) which shows

> "... that an extraordinary degree of comparability exists ... which would be difficult to achieve in public examinations."

8.29 In the context of GCSE and A level, consensus moderation is quite uncommon and there has been alarm in some quarters about the lack of a close interface between the moderation decision-makers and the teachers making the assessments. In the late 1980s Radnor and Shaw undertook a piece of work with a number of local authorities in SW England, developing an assessment and moderation package which sought to be

> "... straightforward, coherent and sensitive to the needs of both teachers and learners, linking notions of public credibility with teacher-based assessment and moderation practices." (Radnor & Shaw, 1995).

8.30 What they developed was firmly rooted in the sorts of principles which some of the CSE boards had sought to operate for many years, but now related to the need to support more authentic assessment practices, and to operate in the context of the emerging GCSE examination. The authors establish a couple of very worthy central principles. The first is the continuity of teachers' assessment activity, directed at a number of purposes, and not easily divisible into simple categories. Moderation has an effect on the whole of this activity. The second is the notion of moderation as a collaboration between insiders (the teachers) and outsiders (external moderators), leading both to decisions and to professional development in assessment. In this respect the very democratic model which they go on to describe merges aspects of agreement trialling and moderation, but has the limitation that it does not provide for the situation where agreement cannot be reached.

8.31 The public credibility of consensus moderation appears to depend on the involvement of an individual who carries the authority of the awarding body, and operates over a number of centres (Radnor & Shaw's outsider). In the early operation of national curriculum assessment at key stage 1 outsiders were provided by local authorities, though it was one of the more peculiar features of the scheme that there were no effective mechanisms for standardising judgements between LEAs (Daugherty, 1995).

8.32 For insights into the relationship between assessment decisions and moderation we have to go back to research into national curriculum assessment. Gipps and others have frequently argued for clarification of the purposes of assessment: if it is for certification or accountability it needs to have an adequate level of reliability. If it is for formative

purposes content and construct validity are more important. There is a possibility of a trade-off between these requirements, especially if the single-minded pursuit of maximal reliability can be abandoned, in favour of a more integrated approach. She argues (Gipps, 1994) that

> "... enhanced validity offered by teacher assessments is gained at a cost to consistency and comparability. Moderation is the process of attempting to *enhance* reliability which... can never be as great as in highly standardised procedures with all pupils taking the same specified tasks."
>
> Then,
>
> "... in line with the professional aspect of teacher assessment, forms of moderation which are based on quality assurance and result in teacher development and enhanced understanding of the subject matter are to be preferred."

Thus, consensus moderation, which involves the discussion of the criteria for assessment decisions, is the most valuable, though very costly.

8.33 Her research here was concerned with teachers' strategies for making assessments in key stage 1, and these provide some insights into the ways in which differences are likely to appear in consensus moderation sessions. She identified three types of strategy for assessment:

- Intuitive: relying on memory; no reference to the statements of attainment; not taking notes; rejecting formal recording; resisting the criterion-referenced approach.
- Evidence Gathering: systematically collecting lots of evidence; making judgements at the end.
- Systematic Planning: planning for teacher assessment; identifying appropriate activities and tasks; using multiple techniques of assessment.

Although this is a simplified representation of her model, it is interesting that the categories all relate to the teachers' approaches to the management of learning, and the first two categories appear to have some overall perceptions of levels that they apply. This suggests that there is first a need to reconcile decisions made on the basis of holistic judgements and those made on the basis of the strict application of the assessment criteria, and then to seek to relate the two together

8.34 Other work by Filer (1993, 1994) discussed the view that teacher assessments must be viewed in the contexts of the learning structures and classrooms in which they take place. This led to the conclusion that a post-hoc quality assurance process cannot achieve comparable assessment of, say, writing in a classroom, since the nature of this activity is determined by individual emphasis on matters such as the use of imaginative ideas or the structure and quality of the writing itself. However, agreement trials should provide a basis for consensus about how these attributes are to be assessed.

8.35 In her discussion of the relationship between assessment and pedagogy, Broadfoot (1996) has provided a similar insight. She discusses the contrast between invisible pedagogies (where teachers' assessment is diffuse, not easily categorised, and does not lead to comparisons between pupils or schools, but does serve the private relationships between the teacher and the taught) and visible pedagogies (where assessments are in relation to visible criteria, capable of supporting comparisons, but which reinforce a teacher-centred learning environment). She discusses evidence from work in primary

classrooms which documents teachers' preference for invisible pedagogies, with their intuitive and idiosyncratic assessment strategies, and their impatience with the categoric assessment structures of the national curriculum, which they see as producing only data for reporting processes. If this tension between the support of learning and the apparatus of order and control is real it suggests that we cannot have seamless teacher assessment activity, and that teachers may tend to reduce highly categorised assessment frameworks to intuitive holistic judgements, perhaps because they are unworkable as they stand.

8.36 The moderation of national curriculum assessment has moved away from the model originally proposed in the Task Group on Assessment and Testing (TGAT) report (for reasons discussed by Daugherty, 1995). The TGAT proposal for group moderation was seen to be too costly and lacking sufficient control of standards; the subsequent debate over the method to be used arrived ultimately at moderation systems which acted as regulatory and administrative processes, rather than a professional one leading to both control and feedback to teachers. At key stage 1 it became a system of moderation by inspection, conducted by each LEA working more or less independently of all other LEAs. At key stage 3 control was to be exercised through a process of audit rather than moderation, designed to deliver dependable results for use in performance tables, seen as a policy imperative. However, the costs of these systems, across so many schools, teachers and pupils were alarming, particularly since the simple procedures, using marks which could be scaled, were not readily available in a criterion-referenced system, especially one as complex as this. In fact, following the Dearing Review of the national curriculum all effective moderation processes were removed. At the same time the credibility of teacher assessments in the national curriculum appeared to be determined by the similarity of the results to those gained by their students on the tests. Burstall (1994) seemed to see it this way, and various evaluations suggest that this was almost certainly the case.

Verification

8.37 As a process, verification has often not had a good press. In its original form, as a mechanism for accreditation and a check on assessment process, it appeared to be unable to cope with the extensive criticism that the actual outcomes of assessments were not comparable from candidate to candidate or centre to centre. The report by Pedreschi *et al* (1994) on management Scottish/NVQs (S/NVQs) airs some anxieties, including concerns about the interpretation of the standards and a lack of consistency in

- assessment between centres and awarding bodies
- defining the roles of those involved in the conduct of the S/NVQs, and their qualifications and experience
- defining what is acceptable and sufficient evidence to demonstrate competence
- the assessment of underpinning knowledge and understanding

8.38 There appeared to be no mechanisms for assuring comparable judgements between awarding bodies. Centres seemed to be very dependent on the decision of external verifiers to whom they looked for advice and support, though it was not clear that they would always act consistently in arriving at assessment decisions. Consequently the external verifier role has steadily increased in complexity, tending to include moderation and guidance as well as verification, and may be close to being almost

impossible to manage within the resources available. These concerns are mirrored in other reports (e.g. Wilmut, 1994a, 1994b, 1995) covering NVQs in the care sector and in business administration, emphasising centres' general dislike of the presentation of the standards, difficulties over sufficiency of evidence, a lack of opportunities for assessors to meet to discuss assessment decisions, and concerns about the competence of some external verifiers. Whilst the standards of external verification have undoubtedly risen over the last few years it remains a difficult task.

8.39 The effectiveness of the verification process is intimately linked with issues concerning the nature and quality of the vocational standards and the varieties of contexts and settings in which NVQs are undertaken. Thus some NVQs are pursued almost exclusively in workplace settings whereas others are mostly delivered by colleges and training establishments in which formal teaching is linked to work placements, using peripatetic assessors. Different organisations make different provisions for the training and support of assessors and internal verifiers and Warmington & Wilmut (2001) suggest that there is some way to go before centres and awarding bodies offering NVQs take consistent views of the roles of internal verifiers (IVs):

> "The integration of internal verification into a quality management framework occurs when organisations perceive and describe an explicit link between internal verification for NVQ and wider issues of quality assurance. Here, the organisation's quality objectives are centrally driven, adhering to planned product, training and inspection standards, for which individuals have explicit responsibility. The quality framework possesses a momentum designed to ensure maintenance of organisational standards, based upon the link between the specification and implementation of organisational standards, their monitoring, evaluation and amendment, leading into the evolution of the strategic plan."

and

> "It is possible to identify a set of baseline conditions that form an irreducible minimum framework within which IVs impact beneficially upon the quality concerns of their organisations, and also feel adequately supported and professionally developed. These may be summarised as
>
> - a coherent accreditation structure for staff operating as IVs
> - NVQ provision incorporated into the organisation's strategic planning for learning and upskilling
> - awarding bodies stipulating the set of tasks that IVs are expected to undertake
> - reference to internal verification duties incorporated into IVs' job specifications
> - organisations specifying the amount of time that IVs are expected to spend on verification duties
> - IVs accorded time allowances or paid increments in recognition of duties undertaken
> - organisations providing adequate resources
> - organisations creating forums where IVs can raise staff training and development issues
> - criteria for the appraisal or evaluation of IVs' performance."

8.40 Particular issues such as that of sufficiency of evidence (a major difficulty in making decisions about the acceptability of candidates' work) have been studied in a number of reports; amongst them are Raggatt & Hevey (1995), Black (1992) and Mitchell & Bartram (1994). There is also the issue of whether a candidate has to 'pass' on every assessment occasion: it is clear from work on reliability of NVQ assessment that a certain amount of compensation goes on, leading to a holistic decision of competence. But transferability is important (is it likely that the candidate, having demonstrated competence in a number of settings, will be able to perform competently in other equivalent contexts?), suggesting that simply repeating assessments in the same context is not going to meet conditions of sufficiency. Mitchell & Bartram see the statements of evidence requirements in S/NVQ standards as the place where sufficiency is defined, and it is noticeable that evidence requirements have become the main focus of some recent specifications, such as those recently published for key skills.

8.41 Concerns about verification in NVQs were mirrored in GNVQ so that Goff & Leimanis (1995) have reported the results of a scrutiny of internal assessment in GNVQ (which was then all of GNVQ assessment) and note, amongst other things,

- differences in quality control practice between awarding bodies
- the need to clarify external and internal verifier roles
- issues of the professional expertise and current backgrounds of external verifiers
- whether external verifiers need to be subject specialists, and their training needs
- greater clarity over sampling – what is sampled and how much is sampled.

Summary

8.42 In relation to the general requirement for quality assurance and control:

- Control over internal assessment can be provided before it starts, as part of its process, or by looking at its outcomes; just about every combination of these approaches has been used. (para 8.1)
- Teachers and students interact very closely with whatever methods of quality assurance and control are implemented. (para 8.3)
- There may be considerable value in an approach to the quality assurance of assessment that regards this as one aspect of the management of all aspects of quality across an institution. (para 8.5)

8.43 In relation to support for teachers:

- Support strategies may provide teachers with materials designed to support their assessment; may provide training in the operation of particular assessments; or may seek to develop teachers' understanding of the principles of assessment. (para 8.7)
- It is possible that too much of the training of teachers has been in the provision of short programmes focused on specific qualifications that do not provide the continuity or depth that is required for real professional development. Yet, high quality provision would be costly and time-consuming. (para 8.10)

8.44 In relation to moderation processes:

- Moderation may be seen as an alerting process through spot-checks or statistical methods or occasional scrutinies when danger signs appear. It may not need to be applied to centres that have met a number of quality criteria. (para 8.6)
- Moderation should be part of a process that gives teachers training, includes feedback to teachers and includes moderator training and standardisation. Moderation should be done without knowledge of teachers' marks or grades and sample sizes for moderation should be adequate to ensure comprehensive judgements. Statistical moderation against written papers as criterion should not be used as a sole method, but should trigger inspection. (para 8.11)
- Statistical moderation is poorly understood; teachers see it as inappropriate to use written papers as a moderating instrument for work of a practical nature. (para 8.13)
- Schools often receive very little feedback on the reasons for changes to marks as a result of moderation, and there is therefore little basis upon which teachers could modify their assessment practice. Decisions made within moderation were seen to take little account of the conditions under which teachers worked and how these affected assessments which they make. (para 8.13)
- Any form of statistical moderation using a reference component consisting of a written paper is open to the criticism that the procedure is statistically illogical. Whilst simple statistical moderation is seen to be cheap and objective it is weak in situations where objectives covered in the school assessment are not covered in the test and where entries are small. (para 8.16 and 8.19)
- Statistical moderation generally involves the scaling of marks whilst preserving the rank order of candidates within a group. It is open to manipulation to maximise results. (para 8.17 and 8.19)
- In moderation by inspection there is little direct evidence on the behaviour of moderators or of the processes that they use; and the design of moderation procedures has generally not drawn on detailed fieldwork with teachers and schools. (para 8.22)
- Problems arise when too wide a range of expectations are attached to a quality control process which is insufficiently resourced. These include moderators being required to focus too much on ensuring reliable assessment and not enough on validity, and being unable to provide the required quality control as well as support for the teachers. Moderators differ in the emphasis which they put on these roles. (para 8.23)
- Consensus moderation has been regarded as a component of a golden era of internal assessment, mainly on the grounds that it promoted professional development as well as achieving the required levels of standardisation. (para 8.25)
- In fact consensus moderation was often done on a goodwill basis with many costs actually being borne by schools. It has been used as a method of operationalising teacher control. (para 8.26)
- A distinction should be drawn between procedures which focus on professional development in assessment (agreement trials) and those which focus on making or ratifying decisions about assessment for a specific purpose (moderation). (para 8.27)

- Forms of moderation which are based on quality assurance and result in teacher development and enhanced understanding of the subject matter are to be preferred, but are very costly. (para 8.32)

8.45 In relation to processes of verification:

- Verification has often failed to cope with the extensive criticism that the actual outcomes of assessments are not comparable from candidate-to-candidate or centre-to-centre. (para 8.37)
- The external verifier role has steadily increased in complexity, tending to include moderation and guidance as well as verification, and may be close to being almost impossible to manage within the resources available. (para 8.38)
- There is a set of baseline conditions that form an irreducible minimum framework within which internal verifiers impact beneficially upon the quality concerns of their organisations, and also feel adequately supported and professionally developed. These may be summarised: as a coherent accreditation structure; the incorporation of assessment and verifications into the organisation's strategic planning for learning and upskilling; awarding bodies stipulating internal verifiers' tasks; specifying the amount of time that they are expected to spend on verification duties. (para 8.39)

9 Internal assessment and lifelong learning

9.1 Perhaps the most persuasive argument for including an internal component in a summative assessment is that it has the capacity to increase the validity of the assessment as a whole. That means that there are domains of students' learning or performance that are beyond the reach of external written assessment; they may be capable of being reported because they have been witnessed by an assessor as they occur, or they may require an extended activity in order to display them, or they may be assessed through group work, or such like. Clearly, the internalising of assessment is not enough by itself – little is achieved if teachers are assessing exactly the same things as an external test or examination other than a hiding of the costs and workloads attributable to the examination. This would not be an adequate justification for the introduction of internal assessment.

9.2 Some commentators allege that some internal assessment for summative purposes has become very stereotyped. This comes from, for example, perceptions that examination coursework is increasingly narrowly focussed and observations that portfolios presented for key skills, GNVQs, VCE and NVQs contain little that can be attributed to student innovation and choice. Whilst much of the evidence is anecdotal, and there are undoubtedly examples of very innovative and student-driven work submitted for assessment by teachers, it may be important to consider whether any degree of stereotyping is likely to inhibit the development of the individual as a lifelong learner.

9.3 Some of the discussions of this stereotyping have been included elsewhere in this review. It is particularly worth noting Black's comments in relation to science coursework and classwork:

> "…the routines and rules set up by the examining groups, together with the lack of training and confidence of teachers in their capacity to guide and assess students' practical investigations, have been turned into a ritual in which training in the component processes dominate and both students and teachers seem to have lost sight of the purpose of collecting scientific evidence"

He cites Duggan and Gott (1996), Hacker & Rowe (1998) and reports similar evidence from Australia (Watson *et al*, 2003).

9.4 Writing about the use of coursework in GCSE science Buchan (1993) suggests that the use of tightly constructed common tasks and procedures tends to squeeze out local context influences, making the internal assessment highly artificial in some cases. Hence, the ways of working demanded of teachers in conducting internal assessments may not be compatible with good classroom practice (and, by implication, the desired learning outcomes for students).

9.5 The present Chief Inspector of Schools appeared to recognise the issues when he commented on the effects of the pursuit of targets

> "… one of the things inspectors find is that an excessive or myopic focus on targets can actually narrow and reduce achievement by crowding out some of the essentials of effective and broadly-based learning. They also find teachers, heads and local authorities for whom targets are now operating more as a threat than a motivator, more as stick than carrot… I have a very real concern that the innovation and reform that we need to see in our schools may be inhibited by an over-concentration on

> targets. It would be an irony indeed if the tool of improvement ended up inhibiting the improvement that is now required." (Bell, 2003)

He emphasises the importance of targets but the implication of his view is that teacher assessments within the national curriculum are being based on too narrow a range of learning.

9.6 Evidence already cited (Weeden & Winter, 1999; Ecclestone & Pryor, 2003) suggests that, within GNVQ and VCE, students evolved safe identities as learners and adopted strategies that would ensure almost minimal compliance with assessment requirements in order to get the qualification that they required. This also emerged strongly in an unreported review of key skills portfolios, conducted as part of the evaluation of the key skills qualification pilot where it was difficult in some cases to determine each student's individual decision-making within a class group of portfolios, all of which contained almost identical task sheets, assignments and exercises.

9.7 Ecclestone and Pryor (2003) have constructed a persuasive discussion of the interplay between assessment practices and the development of learners' identities and note that both formative and summative purposes have an essential role in this, exerting ever stronger influences as learners progress through education developing learning dispositions that relate to lifelong learning and the development of learning careers. This concept of a learning career may be viewed in relation to economic, occupational and social, as well as educational factors, and may be related very closely to Lave and Wegner's (1991) presentation of learning as both socially constructed and context specific and therefore influential in exposing learners to new influences and situations that change dispositions. Thus, the experience with GNVQ or any other programme that contains an assessment regime, interacts with external factors and existing dispositions to construct students' images of ability, acceptable teaching and engagement with assessment activities. This is a step in the construction of students' assessment careers and the socialisations involved in the development of these careers have implications for the ways in which students move between qualifications as they progress through lifelong learning.

9.8 The link between assessment processes and learning management is also made in the discussion of assessment for learning with the 10 principles produced by the Assessment Reform Group (2002b) including

> "Principle 9: Assessment for learning should develop learners' capacity for self-assessment so that they can become reflective and self-managing
>
> Independent learners have the ability to seek out and gain new skills, new knowledge and new understandings. They are able to engage in self-reflection and to identify the next steps in their learning. Teachers should equip learners with the desire and the capacity to do this for themselves through developing the skills of self-assessment."

9.9 In that case we need to be sure that we are creating the basis upon which such development can take place. Should students never be required to take decisions related to a piece of work, evaluate findings, reflect on processes and their own performance, set targets or interact with others, they may not be equipped to develop learning careers to an extent that will cause them to function as lifelong learners. It may also be the case that they can never be effective scientists, historians, engineers, artists, builders or managers without these skills. Whilst it is correct to lay emphasis on the acquisition of skills of communication and the use of number these do not represent the whole of what

is needed for lifelong learning, nor the whole of what is needed for employability. Thus, with the 14-19 Review placing considerable emphasis on the development of generic skills within programmes it may be appropriate to consider how these are to be integrated into the whole learning programme for an individual and how far approaches to assessment (and particularly internal assessment) support or inhibit their development.

9.10 This opens up the resonance with a variety of initiatives on records of achievement, action planning, progress file and personal development planning, all of which have explored aspects of the decisions made about learning between students and their teachers or assessors. Whilst these initiatives have often developed teacher (and student) expertise very considerably, what has remained has been the documentation; some of this has continued to provide very effective underpinning for desired educational processes, but some has lost its purpose under pressure from later demands.

9.11 One important current programme is the *Assessment is for Learning* initiative in Scotland that aims to provide a streamlined and coherent system of assessment designed to ensure that parents, teachers and other professionals have the feedback they need on students' learning and development needs. The programme (Learning and Teaching for Scotland, 2004) aims to

- develop one unified system of recording and reporting, the Personal Learning Plan (PLP), which will bring together the current PLP, Progress File, transition records and Individualised Educational Programmes (IEPs)
- bring together current arrangements for assessment, including the Assessment of Achievement Programme, National Tests and the annual 5-14 Survey of Attainment
- provide extensive staff development and support through its project-based approach.

This is seen to have the benefits of

- better feedback for pupils leading to improved achievement
- a simplified system and support for teachers leading to a reduction in workload
- clearer information for parents.

Summary

9.12 Linking assessment and the development of the lifelong learner:

- Little is achieved if teachers are assessing exactly the same things as an external test or examination other than a hiding of the costs and workloads attributable to the examination. This would not be an adequate justification for the introduction of internal assessment. (para 9.1)
- Some commentators allege that some internal assessment for summative purposes has become very stereotyped. (para 9.2)
- Some students have evolved safe identities as learners and adopted strategies that ensure almost minimal compliance with assessment requirements in order to get the qualification that they require. (para 9.6)
- Both formative and summative purposes have an essential role in exerting ever stronger influences as learners progress through education developing learning

dispositions that relate to lifelong learning and the development of learning careers. (para 9.7)

- Assessment for learning should develop learners' capacity for self-assessment so that they can become reflective and self-managing. (para 9.8)
- Should students never be required to take decisions related to a piece of work, evaluate findings, reflect on processes and their own performance, set targets or interact with others they may not be equipped to develop learning careers to an extent that will cause them to function as lifelong learners. (para 9.9)

10 Internal assessment in the national curriculum

10.1 Whilst this review is focused on the 14-19 curriculum a major slice of the work on internal assessment has been undertaken in relation to national curriculum assessment and there have been many references to this earlier in this review. Although we should not assume that models for internal assessment can be transported uncritically from that experience to operate within qualifications in the 14-19 sector it is many of the same teachers who will be involved in both, and many studies have pointed to the exchange of expertise between the two areas. This short section provides a background to some of the points already made.

10.2 Although the TGAT report saw teacher assessment as having a central role in national curriculum assessment (on the grounds that the national curriculum was extremely complex and beyond the scope of any practical external testing) it was not the primary one; it said that "… written test responses might be supplemented by teacher observation of skills made in a systematic way". This put teacher assessment firmly in its place and it has still not emerged from the shadows; Daugherty (1995) believed that this was because

- policy makers did not see it as an important component of the system, and their agenda did not link it to formative and diagnostic aspects of assessment
- the development timetable and funding focused around test development and implementation.

10.3 Relatively limited resources were allocated to teacher assessment in the national curriculum, support and information were provided late and early public statements did not give it much prominence. Public discussion tended to focus on the problems that it raised rather than the possibilities that it offered and the political focus was on measuring attainment, stimulating competition and the local and national accountability of schools (Torrance, 1995). These points were not lost on schools, and teachers undertook what were enormously complex assessment tasks largely without central guidance and support, with a poorly articulated purpose for what they were doing, and no clear relationship in the core subjects between teacher assessment and the tests.

10.4 Some of the issues are the same as those encountered in teachers' assessments within GCSE, Graded Assessment, Records of Achievement and the like. Notions of what counts as assessment have been too narrow, the focus has been on assessment rather than learning, and teachers have had limited opportunities to develop assessment methods that support and serve their curriculum goals. Teachers in the primary sector were particularly disadvantaged in the early stages, being faced with a model that drew heavily on secondary experience, which they did not share. McCallum *et al* (1995) describe the emergence of a variety of pragmatic approaches in which infant teachers attempted to develop their understanding of criterion-referenced assessment within their general practice and philosophy of primary education. Particular tensions surrounded their need to take account of 'the whole child' (and therefore to over-ride criteria with more global judgements) and the risk that they would import externally devised assessment resources and turn these into mini-tests in order to satisfy the system requirements.

10.5 Some of the issues were illustrated in a number of other studies that demonstrate the adaptations that teachers have made to their programmes of study in order to ensure that they could make assessments that they are able to justify, whilst retaining what they regard as valid and meaningful classroom learning (for example, Gipps, 1995; Hall *et al*, 1997; Emery *et al* 1998a, 1998b). In an effort to support teacher assessment, SCAA and later QCA provided a range of exemplar and supporting materials, together with extended proposals for ways in which departments or other groups of teachers might develop internal quality assurance using a mix of internal and external examples for discussion in round-table meetings, and insertion into school portfolios. These arrangements have not, for the most part, survived.

10.6 At the same time, in a theoretical discussion of the concepts of validity, dependability and reliability of national curriculum assessment, Wiliam (1993) takes the view that it is running ahead of theoretical frameworks available to support it, and it seems to be the case that teachers, faced with competing demands, conflate the formative and summative functions of assessment (Ecclestone & Hall, 1999).

10.7 Firestone (1998) conducted a comparative review of national curriculum assessment developments in England and parallel developments in Vermont, USA. In commenting on the TGAT proposals he observed that

> "Formalised formative assessment is a crucial but ambivalent idea. It re-establishes teachers as professional decision-makers in spite of conservative attacks on them, but it suggests that, without standardized assessment procedures, teachers lack a rational basis for good instructional decisions."

He noted that the incorporation of assessment into programmes of work was not a feature of the early days of national curriculum assessment, but that this followed the simplification of the national curriculum, as a result of the Dearing Review. He appears to suggest that teachers need to place some value on assessments before they will make these linkages:

> "Teachers... will not use assessments to support instruction unless they have such high face validity that [they] want to use them."

10.8 In discussions of the ways in which teachers make assessments within the national curriculum (particularly in relation to the idea of ensuring the 'best fit' of evidence to the level descriptions) Gipps *et al* (1998b) suggested that teacher judgements are not wholly consistent from school to school: they identified several different strategies in relation to the idea of judging the best fit of the work done by a student in a variety of contexts to each attainment target. Some teachers made intuitive general best fit judgements, others judged more globally in relation to work in portfolios, others used some type of mechanistic process-based on splitting the level descriptions, while others identified the presence of key aspects of level descriptions. Internal standardisation processes were thought useful and effective (though these, and the use of school portfolios, may have died away somewhat). They did suggest that the combination of exemplification materials and group moderation is a desirable process for arriving at common standards, but also reported mixed opinions about the external exemplification materials, though these were more welcomed by primary than by secondary teachers.

10.9 More recently the QCA Foundation Stage Profile Handbook for 2003 makes provision for moderation meetings and visits, operated by LEAs. This represents a return to procedures that operated in the early days of the national curriculum but which were

abandoned on cost grounds. At the same time a new strategy for primary schools (DfES, 2003) makes provision for tests and tasks to be used as components of the teacher assessment process.

Summary

10.10 Additional evidence from experience with national curriculum teacher assessment includes:

- Teacher assessment continues to operate in the shadow of testing because policy makers have not seen it as an important component of the system and their agenda did not link it to formative and diagnostic aspects of assessment. (para 10.2)
- In the early days of national curriculum assessment teachers undertook the enormously complex assessment tasks largely without central guidance and support and with a poorly articulated purpose for what they were doing. (para 10.3)
- A range of exemplar and supporting materials was provided, together with extended proposals for ways in which departments or other groups of teachers might develop internal quality assurance. These arrangements have not, for the most part, survived although the materials still exist and are valued. (para 10.5)
- National curriculum assessment is running ahead of theoretical frameworks available to support it. Teachers, faced with competing demands, conflate the formative and summative functions of assessment. (para 10.6)
- The combination of exemplification materials and group moderation is a desirable process for arriving at common standards. (para 10.8)

11 Some particular internal assessment strategies

Introduction

11.1 Elsewhere references have been made to particular aspects or methods of assessment, three of which deserve rather more discussion. They are grouped together in this section for convenience but they are not related in any other way.

Portfolios and performance assessment

11.2 Within vocational qualifications the principal basis for assessment continues to be through portfolios; these contain evidence assembled by students in order to demonstrate competence. This is designed to place students in decision-making roles that involve them in understanding and interpreting the specification for the qualification, and deciding what evidence it will be appropriate to incorporate in the portfolio. Of course, it doesn't always work like that, and assessors normally provide a considerable amount of guidance and, in some cases, exert almost total control over the evidence-assembling process, often using standard tasks.

11.3 The use of portfolios raises the issue of the interactions between learning development and assessment to the point that some writers distinguish the learning portfolio from the assessment portfolio. This is because the processes by which students manage the compilation of the portfolio are themselves key aspects of learning and the management of learning. Thus, the creation of a portfolio potentially entails the management of a work programme, the creation, evaluation, choice and presentation of items to be included in the portfolio and processes of review and decision-making that relate very closely to some aspects of formative assessment as well as to the key skill of managing own learning and performance.

11.4 The process of review, that may be incorporated into learning portfolio management, has attracted considerable recent attention in higher education in the context of personal development planning (see, for example, Moon, 2000), and Klenowski (2003) has suggested, on the basis of a study of portfolios in an HE programme, that

> "The learning portfolio appeared to encourage the processes of learning and reflection but also offered a way of recording and structuring those processes."

11.5 Because the management of the portfolio is always likely to involve a teacher, assessor or supervisor in a centre it may become a key feature of the conduct of internal assessment. It may be regarded simply as a collection of evidence for assessment for summative purposes ('evidence' is the word widely used to describe portfolio contents), but it may also provide a basis for linking reflection with changes in learning practice for the individual. Klenowski again:

> "What has become apparent in the use of a portfolio for learning purposes is the need to shift the emphasis from the collection of evidence to a focus on the analysis and integration of learning."

As part of the learning management aspect of portfolio assembly and maintenance she also suggests that students should share in the framing of the criteria by which assignments and other portfolio components are assessed.

11.6 Portfolios are not unique to vocational qualifications, nor to this country. Koretz (1998) has described four examples of portfolio assessment in the United States. These were

introduced in reaction to the narrowing of the curriculum following the almost exclusive use of multiple choice testing. It involves the production of a portfolio, but under varying conditions of external control and with varying amounts of externally standardised content. Marking is generally not by the students' own teachers, but by teachers employed and trained for the process. It is thus an instrument of external assessment in a very direct way.

11.7 Koretz looked at the validity and reliability of the portfolio assessment outcomes. Generally the levels of marking consistency were low, although they were improved with experience and refinement of the assessment criteria, though much more in some subjects than others. Problems of the differences in domain sampling between students provided by the chosen pieces of work made it difficult to achieve consistent scoring. The evidence for the validity of the assessments was limited though varying between the examples. The author concluded that the portfolio assessments are not of a sufficient quality to enable comparisons to be made between students for selection or between schools for accountability, though the effects on improved instruction may be considerable.

11.8 Another view of some of the same portfolio assessment in the United States (in Kentucky and Vermont) has been provided by Stecher (1998), who has discussed the benefits that are seen to accrue from the incorporation of more authentic assessment, the effects on teachers and schools, and the outcomes. The conclusions are interesting.

- Portfolio assessment places substantial new demands on teachers and schools and there is a need for additional training for preparation and marking.
- There are seen to be benefits in terms of the positive effects on teaching practice and on learning, and in acting as a focus for discussions amongst teachers about learning. The benefits are seen to outweigh the burdens, though this may change if workloads experienced in the early years of the schemes are not seen to diminish.
- Moreover, the benefits must be seen in terms of increased levels of student achievement – teachers see this as a crucial factor.

However, these portfolio assessments operate as part of a high-stakes assessment and it is not clear whether teachers' tolerance of the workloads would be as great if the issues of accountability were not so prominent. The portfolios are designed to broaden the curriculum and to provide a more authentic approach to assessment. This may be limited by the use of explicit scoring criteria, which are more precisely stated than the curriculum itself; teachers may interpret these as the curriculum goals, with potentially damaging effects. In other words, something as flexible as portfolio assessment may be very limiting if not operated in the context of clearly articulated curriculum goals, which operate as the principal point of reference.

11.9 Similar use of portfolios is reported in Canada, in work in elementary schools (Anderson & Bachor, 1998) and in South Africa (Johnson, 1998), in primary schools. This latter work broadens the discussion to consider the use of performance assessment, and the major difficulties involved in meeting the requirements for teachers to adjust to the requirements of operating this approach to assessment. This is not simply a matter of more training, though this is needed. Johnson discusses the Maxwell & Cumming (1998) review of the process of adjustment of teachers to the use of performance-based assessment. They say that the sort of paradigm shift that is involved is very difficult and

time-consuming to achieve, especially where large numbers of people are involved. For teachers it requires

> "... a process of individual intellectual struggle on the part of all teachers in order for the new paradigm to overthrow the old. For this to be successful, teachers need to perceive some purpose or advantage in engaging in the struggle."

11.10 Harlen (2004) has also discussed the assessment of portfolios as part of her Evidence for Policy and Practice Information (EPPI) study on the validity and reliability of various forms internal assessment. Because of the constraints of the EPPI methodology she only included three studies directly related to portfolios (one was an earlier paper by Koretz on the Vermont programme), none of which discuss the portfolio as a learning device or as a means of assessing the management of learning. However, there was a strong suggestion from the work by Koretz and by another study of the use of portfolios in primary grades in Texas (Shapley & Bush, 1999) that teachers' assessments could only be made acceptably valid and reliable if the tasks and the criteria by which they were assessed were externally supplied and closely controlled. As Harlen points out, this does create a tension with the need for learners to take some control over the portfolio and for teachers to have some ownership of the assessment process, and it may be unwise to base conclusions about the use of portfolios on such a narrow evidence base.

Assessing practical, oral and other performance work

11.11 Wood (1991) discussed the assessment of practical work, chiefly in science. He noted that the one-off practical examination is a very weak assessment instrument. The rationale for practical assessment is strong, since it has the potential to cover many skills that cannot be assessed in written tests, although some of what is assessed in a practical examination may be very close to the cognitive areas covered by written papers (Hoste, 1982). Early work on A level Chemistry by Wood & Ferguson (1975) suggested that the practical examination was a less adequate measure than teachers' assessments over a number of experiments, and that the most suitable way of assessing practical work (with its complex mix of skills) might be through a combination of written tests (for the cognitive skills) and continuous assessment in the laboratory (for psycho-motor skills).

11.12 Quite a lot of work on the assessment of practical skills was done in the 1980s, often as part of the work on Graded Assessment. Typical was the work of Lock & Ferriman (1989) in connection with the Oxford Certificate of Educational Achievement (OCEA), on graded schemes in languages, related to the assessment of oral skills, and a range of experience was developed in the work of the Assessment of Performance Unit (APU) where Gott (1986) discussed the links between methods of assessment and desirable ways of science learning, including the role of the teacher in supplying appropriate tasks and in providing appropriate support during the practical work.

11.13 The assessment of practical work often involves teachers in observing students working in groups, and there Wood reported some evidence of the difficulties involved in this when there are many students to be watched and many constructs on which to report. The problem is the same as that in the national curriculum, where the demands on teachers have now been reduced by seeking more holistic judgements. There seems to be little evidence of the effects of this on validity and reliability. The other problem is that of student-student interaction, but that too seems little researched in relation to

assessment judgements. However, Wood does discuss elaborative procedures, where assessments are made as a result of probing, observation and clarification over a period and over a number of activities, and Wood & Power (1987) have argued that this is, perhaps, the only satisfactory way of establishing a decision regarding competence which can be done in a way which establishes the limits of performance.

11.14 Having described the main characteristics of assessment for learning Black (2002) reiterates the conclusions drawn by Black and Wiliam (1998): that assessment methods used by teachers are not effective in promoting good learning, that marking and grading tend to emphasise competition rather than personal improvement, and that assessment feedback often has a negative impact, particularly on pupils with low attainments who are led to believe they lack ability and are not able to learn. He recognises the needs to continue to serve the different purposes of assessment, to achieve optimum reliability and validity so as to command public confidence, and to strengthen teachers' assessments in general. He suggests that one way forward would be to go back to project-based assessments, perhaps as part of a portfolio approach and to ensure that there is a greater diversity of assessments used for summative purposes.

Modular approaches and internal assessment

11.15 The greater development of modular or unit-based approaches and syllabuses in the 1980s has from time to time involved a greater use of internal assessment. These programmes have often suffered low esteem; Howieson (1993) pointed to the problems with the modular provision in Scotland, where low esteem of the modular provision introduced by the Scottish Vocational Education Council (SCOTVEC) was associated with the use of internal assessment. But it also had a lot to do with the nature of the take-up and the lack of recognition by Higher Education. She says:

> "… if parity is to be achieved any new system must be a unified one in which academic and vocational elements are integrated within a single award and that there needs to be a greater emphasis on external assessment and the introduction of graded awards."

11.16 There are perceived to be problems of learning fragmentation within modular approaches, leading to the requirement for a form of synoptic assessment to be included. In a very useful review Patrick contrasts the different requirements of synoptic assessment in a range of A level programmes (Patrick, 2003). Synoptic assessment is required on validity grounds and is normally in the form of an external assessment taken at the end of a course, though there is the possibility that an element of internal assessment may be involved in some cases. So, for example, in business studies and economics, synoptic assessment might be conducted through 'internal assessment requiring candidates to apply knowledge, understanding and skills learned in other parts of the course, e.g. a project based on experience of work'. In practice current examinations do not use this option, but two of the awarding bodies do use pre-release case study material that is used as the basis for a written examination. Two physical education examinations include coursework in synoptic components. In psychology one awarding body requires a research report that is externally marked. Patrick mentions further examples of pre-released material in some geography and English literature examinations.

11.17 It seems likely that internally assessed synoptic assessment will be undertaken under tight conditions of external control, and that it will include the assessment of a wider

range of objectives or higher skills than are to be found in many internal assessments. There may be opportunities here for extended pieces of work over a period of time, incorporating elements from many parts of a 2-year programme, and allowing the assessment of this wider range of learning. However, its special status within the whole qualification will probably mean that it will be specified, conducted and controlled externally rather than by a teacher.

Summary

11.18 In relation to portfolios and performance assessment

- Portfolios contain evidence assembled by students in order to demonstrate competence and are designed to place students in decision-making roles that involve them in understanding and interpreting the specification for the qualification, and deciding what evidence it will be appropriate to incorporate in the portfolio. In practice assessors normally provide a considerable amount of guidance and sometimes exert almost total control over the evidence-assembling process. (para 11.2)
- The processes by which students manage the compilation of the portfolio are themselves key aspects of learning and the management of learning. The creation of a portfolio potentially entails the management of a work programme, the creation, evaluation, choice and presentation of items to be included in the portfolio and processes of review and decision-making that relate very closely to some aspects of formative assessment. (para 11.3)
- There is a need to shift the emphasis from the collection of evidence in a portfolio to a focus on the analysis and integration of learning. Students should share in the framing of the criteria by which assignments and other portfolio components are assessed. (para 11.5)
- In some situations the marking of portfolios is not by the students' own teachers, but by teachers employed and trained for the process, making them an instrument of external assessment in a very direct way. (para 11.6)
- Some portfolio assessments are not of a sufficient quality to enable comparisons to be made between students for selection or between schools for accountability, though the effects on improved instruction may be considerable. (para 11.7)
- Portfolio assessment places substantial new demands on teachers and schools and there is a need for additional training for preparation and marking. The intention that portfolios should broaden the curriculum and provide a more authentic approach to assessment may be limited by the use of explicit scoring criteria that teachers may interpret as the curriculum goals. In other words, something as flexible as portfolio assessment may be very limiting if not operated in the context of clearly articulated curriculum goals, which operate as the principal point of reference. (para 11.8)
- Whilst it had been said that teachers' assessments of portfolios could only be made acceptably valid and reliable if the tasks and the criteria by which they were assessed were externally supplied and closely controlled, this does create a tension with the need for learners to take some control over the portfolio and for teachers to have some ownership of the assessment process. (para 11.10)

11.19 In relation to assessing practical, oral and other performance work

- The one-off practical examination is said to be a very weak assessment instrument and some of what is assessed in a practical examination may be very close to the cognitive areas covered by written papers. A practical examination is a less adequate measure than teachers' assessments over a number of experiments. (para 11.11)
- The assessment of practical work often involves teachers in observing students working in groups and there are difficulties involved in this when there are many students to be watched and many constructs on which to report. The problem of student-student interaction is little researched in relation to assessment judgements. (para 11.13)

11.20 In relation to modular approaches

- It is possible that internally assessed synoptic assessment will be undertaken under tight conditions of external control, and that it will include the assessment of a wider range of objectives or higher skills than are to be found in many internal assessments. There may be opportunities here for extended pieces of work over a period of time. (para 11.17)

12 Making progress: a discussion

Introduction

12.1 This review has attempted to represent the diversity of internal assessment activity in relation to summative purposes over the last 50 years, and some of the factors that have affected its development. I have generally focused on the 14-19 sector and the qualifications that operate in that sector in the UK, but have tried to draw on relevant ideas and evidence from elsewhere in education and training. This section seeks to draw this together by attempting to address four questions:

- What general principles, grounded in appropriate frameworks of assessment, learning and progression, might underpin models for teacher assessment for summative purposes?
- What are the threats and risks (in relation to reliability, validity, public confidence etc) associated with moving towards a far greater reliance upon teacher assessment in general qualifications?
- How feasible is it for teacher assessment to support both formative and summative purposes?
- How, if at all, does teacher assessment address the issue of the overall burden of assessment on teachers, students, awarding bodies, etc.?

General principles and models

12.2 Are we close to a single model of internal assessment for summative purposes? The answer to this must be 'no' and this may be true for the foreseeable future. The conditions that operate in relation to the national curriculum are not the same as those that operate 14-19, where the stakes are different. The approaches that are used for occupational and some parts of vocational provision have evolved differently. However, there is a large amount that is held in common between the different areas, and it is appropriate to try to draw ideas and approaches from all areas and, to some extent, from overseas experience, in support of a discussion of where to go next.

12.3 If we are not close to a single model are we able to articulate purposes and principles for teacher assessment that can be applied in a variety of situations? The answer to this may be a qualified 'yes' since we are probably better able to articulate the key issues that must be addressed and we can provide some indications of the answers to these. They are, in fact, derived from the broad elements of the taxonomy shown in Section 3, namely

- the overall assessment design that is to incorporate an element of internal assessment in terms of its general structure, the domains to be assessed, how the internal assessment relates to the learning programme and how roles are distributed in relation to it
- how tasks are to be originated and specified, how many there will be and how long they will last, the form that they will take and the evidence that they will generate
- the assessment processes and the basis upon which they will be conducted, how teachers will make and record judgements and how their expertise will be developed so that they can do this effectively

- the methods that will be used for quality assurance, who manages and controls these and how the tasks are distributed internally and externally.

12.4 How far can all of this be grounded in appropriate frameworks for assessment, learning and progression? Perhaps the answer to this is that we first need to draw on the sorts of principles that have been developed by the Assessment Reform Group. It is difficult to envisage violating principles of that type as a result of the imposition of summative teacher assessment; the risks to learning would be too great. However, in the current climate, the demand for dependable assessment leading to credible and valued qualifications that enable progression means that requirements for valid and reliable assessment must be met, although the basis for judging what is reliable and what is valid may need to be much more sophisticated than it is at present.

12.5 Additionally, the biggest anxiety about learning would be that it might not provide individuals with the skills and knowledge that they need for progression. These are not just skills and knowledge related to subjects or occupations but those related to learning itself: life skills, learning to learn skills, core skills, key skills or basic skills. All of these represent commitments to future development rather than immediate achievement. Approaches to learning that are strongly directed towards the achievement of a qualification, to the exclusion of other concerns, may squeeze these skills out, since they are of long-term rather than immediate concern. The narrowing of learning (to which summative assessment requirements may contribute) may be at odds with the development of the individual as a lifelong learner.

Threats and risks from more summative teacher assessment

12.6 Apart from issues of workload and professional development (discussed below) the first and greatest difficulty is concerned with perceptions of the reliability of teacher assessment. It has been widely assumed for the last 50 years that the written examination was the primary means of getting dependable assessment and that internal components were only needed when validity demanded it. There have been exceptions to this view but they have not managed to invert this priority. The considerable risk is that an increased commitment to teacher assessment for summative purposes will be regarded as a cause of lowered standards or loss of rigour and be cut back to a level where it is thought to be safe.

12.7 There may be four ways of reducing (but certainly not eliminating) this risk. They are

- mounting a concerted campaign to give internal assessment at least an equal status with written examinations
- developing teacher expertise in assessment to a point where there can be an expectation of reliable and valid judgements
- developing methods that ensure unambiguous specification of the basis for the internal assessment
- providing strong and appropriate quality assurance and control structures.

Each of these strategies will have a limited effect by itself, but may together be effective over a long period.

12.8 The difficulties associated with a campaign are obvious, and it will need to be seen as more of a ling-term change in public rhetoric than a singe publicity exercise. Without it, however, it will be difficult to persuade teachers that the changes demanded of their

practice are worthwhile and the public that there has not been a dilution of educational standards. Similarly, the development of teacher expertise cannot be seen as a short-term commitment and it will need to be accompanied by clear evidence that the effort required is professionally worthwhile. There are precedents for this but it will not be achieved without considerable resourcing over a long period.

12.9 The effects of the campaign would be somewhat enhanced if it were backed by steps that reduced the stakes associated with summative assessment. It is difficult to see how this can be done directly for students but a reduction in the use of examination results as a basis for judging institutional success is possible and would make it easier to introduce a greater degree of teacher control over assessment for summative purposes.

12.10 The specification of the internal assessment is more easily addressed using the technical expertise that we already possess perhaps supplemented by further research. This may involve developing frameworks within which teachers can develop and use their own assessment tasks, the provision of exemplar tasks or banks of tasks from which they can select or the external specification of tasks that they must incorporate. Which of these is most appropriate will depend on perceptions of their expertise and the risks to pedagogy that may result from too much external prescription.

12.11 Quality assurance is an area where there is also some scope for innovation that focuses on two issues:

- the use of methods that integrate well with the development of teacher expertise in the management and conduct of assessment
- approaches that can integrate moderation and verification across qualifications and into wider institutional systems for quality assurance.

In a climate where we are promoting the relationship between assessment and learning it seems logical to look at the mechanisms for quality assuring assessment as one part of wider mechanisms for quality assuring teaching, staff development, resources, management and other activities that happen in the institution. Structures for doing this have a number of important features: they treat quality as a whole-institution provision; they manage it in a systematic way; they establish the relationships between different aspects of the provision; they allocate responsibilities and devise quality assurance methods on an institution-wide basis; they monitor delivery and provide mechanisms for information collection and processing; they build in feedback loops that enable action to be taken in a systematic way.

12.12 This approach clearly offers something that separate inputs from a number of awarding bodies cannot: it is systemic (addressing the environmental issues that affect the assessment process) and it is coherent (allowing the transfer of expertise through the institution). It shifts some degree of control over the quality assurance of assessment away from awarding bodies, but explicitly provides links designed to ensure professional development. It arguably offers the possibility of giving more experienced and expert teachers and assessors more control over, for example, the choice of internal assessment tasks, as well as over the daily conduct of teaching and learning. However, to be effective it needs expert external inputs from moderators and verifiers, as at present.

Supporting formative and summative purposes

12.13 The earlier discussion of principles suggests conditions under which formative and summative purposes may be supported together. Much contemporary discussion focuses on the resolution of the tension between these two assessment purposes; if this resolution were not seen to be appropriate then either the proposition that teachers should extend their role in assessing for summative purposes falls or assessment for summative purposes would develop as a self-contained provision.

12.14 Resolution of the tension appears to require that tasks be curriculum-embedded and probably that there should be many small ones rather than a few large ones. In order for requirements for dependable assessment to be met we may need to know more about teachers' assessment decision-making processes, how they move between specific and global judgements and what influences these decisions, and whether the aggregation of many small assessments ultimately leads to a more reliable overall result than assessing a larger task. The issues are identified in the taxonomy; we need to understand better the effects of

- the types of tasks that will be used
- the forms of evidence that will be assembled
- whether assessment is of mastery or is compensatory
- the use of analytic or holistic judgements
- the nature of the record of the assessment that is produced.

In addressing these issues it seems appropriate to call upon experience with vocational and occupational qualifications in which assessments are already made in relation to a large number of embedded tasks, using a variety of strategies. In determining the appropriateness of various approaches and strategies we will need first to decide the terms in which successful incorporation of tasks for summative purposes are to be defined: what, for example, will be acceptable levels of task reliability and what levels will we demand for the aggregated assessment?

12.15 Returning to the question of risks: a further and important risk attached to the wider use of teacher assessment for summative purposes is that its implementation may not be accompanied by the paradigm shift that is required in order that it might work. Some of this risk is associated with a failure to change the attitudes associated with teacher assessment, some with the problems that may arise from an over-mechanistic approach to the conduct of the assessment, and some from a failure to develop and identify clear advantages in making the change. This last aspect of risk is closely linked with the development of ways in which teachers and students interact to use assessment for the improvement of learning through, for example, the incorporation of processes of reflection and review in internally assessed components in all types of qualifications. There have to be clear learning gains and it will be important that students see the value in probing, analysing, evaluating, discussing and reviewing what they have done, and developing learning plans in the light of this, thus exploiting the possibilities that curriculum-based tasks provide and that are unavailable through external examinations.

The burden of assessment

12.16 Internal assessment has been seen as a way of reducing the examination workload; this has become considerable for both students and their institutions, with associated high

costs. Coursework demands for public examinations have also been very high for students and portfolio compilation is known to make very heavy demands on students working on GNVQ, VCE, key skills and NVQ qualifications. Steps to reduce this pressure (as, for example, in enabling students to offer proxies or work done in other connections) may appear to debase the value of the qualification and attempts to spread coursework over a longer period may be inhibited by modular structures and could increase the pressure and effects on learning.

12.17 Unless some specific steps are taken to avoid it, past experience suggests that an increase in internal assessment for summative purposes will increase overall workloads for teachers and students. The suggestion that internal assessment for summative purposes will draw on work done as part of a learning programme, and therefore not represent a workload increase, may have some justification, but it would be naïve to suppose that this could be achieved without significant additional resourcing and external support. This issue can only be addressed in the context of a much wider discussion that involves

- finding conditions under which effective formative assessment is compatible with reliable summative assessment; that is, that they can co-exist without one damaging the other
- understanding the extent to which teachers are already conducting assessment that is suitable for summative purposes
- determining the amount and type of support for teachers that is needed in order that they can be sufficiently effective assessors for summative purposes
- discovering whether pedagogy can survive the imposition of summative assessment and student and teacher behaviour can be primarily focused on learning
- whether assessment tasks that are controlled by teachers will tend to increase the volume of formal testing at the expense of more broadly based activities.

12.18 The development of teacher expertise in this area seems unlikely to be a short-term commitment through one-off training courses nor likely to be of long-term benefit if it is focused on the requirements of particular qualifications or merely shifts unresolved assessment problems from the awarding bodies to institutions. QCA has accurately predicted the difficulties that exist (QCA, 2003):

> "The current volume and forms of external assessment exist largely because of previous concerns about the lack of public credibility of internal (teacher-based) assessment. It will be essential to ensure that a shift from external assessment to teacher-based assessment does not simply relocate any perceived assessment problems with teachers, many of whom may not necessarily be predisposed to take on the task. A major staff development and support programme will be required if teachers are to take on an enhanced assessment role."

In fact it is possible that the starting point should be the development of expertise in formative assessment that leads into the treatment of assessment for summative purposes rather than looking for beneficial backwash on formative practice coming from the provision of training in summative methods. Improving practice in assessment to the point where teachers are seen to be the source of dependable summative assessment will be a long haul that may need to be accompanied by changes in rhetoric about where 'standards' and 'rigour' are vested.

12.19 We can see an integration between issues of professional development and quality assurance that suggest that we are likely to need

- a treatment of assessment as one part of a wider provision of quality in learning
- the development of institution-based structures for managing and delivering quality
- the provision of support for these developments within institutions and in an integrated fashion
- the development of methods for monitoring the existence and operation of institution-based quality management systems
- the eventual accreditation of institutions to manage assessment alongside other provisions, under a process of light monitoring.

References

ADAMS, R.M. & WILMUT, J. (1982) A measure of the weights of examinations components, and scaling to adjust them. *The Statistician* 30. 263-9

AKYEAMPONG, A. & MURPHY, R. (1997) Introducing Continuous Assessment: Challenges and Changes. *Mauritius Examinations Bulletin* 7. 14-22

ALLEN, R. & TRAVERS, E.J. (1995) *Multiple regression analysis of overall positions and levels of achievement.* Brisbane: Board of Secondary School Studies

ANDERSON, J.O. & BACHOR, D.G. (1998) A Canadian Perspective on Portfolio Use in Student Assessment. *Assessment in Education* 5.3. 353-379

ARCHER, J. & McCARTHY, B. (1988) Personal biases in student assessment. *Educational Research.* 30.2 142-145

ASSESSMENT REFORM GROUP (2002a) *Testing, Motivation and Learning.* Cambridge: Assessment Reform Group

ASSESSMENT REFORM GROUP (2002b) *Assessment for Learning: 10 Principles.* www.assessment-reform-group.org.uk/CIE3.pdf

ASSOCIATED EXAMINING BOARD (1981) *Combining Teacher Assessment with Examining Board Assessment.* Report of a seminar. Aldershot: AEB

BELL, D. (2003) *Reporting for England.* Speech by Her Majesty's Chief Inspector of Schools to the City of York Council's annual education conference, 28 February 2003

BLACK, H. (1992) Sufficiency of Evidence: what might be fair and defensible. *Competence & Assessment* 20. 3-10

BLACK, P.J. (1990) 'Social and Educational Imperatives for Changing Examinations'. in LIUTJEN, A.J.M. *Issues in Public Examinations* Proceedings of the 1990 Conference of the International Association for Educational Assessment. Utrecht: Lemma

BLACK, P. & WILIAM, D. (1998) Assessment and Classroom Learning *Assessment in Education* 5.1. 7-74

BLACK, P. (2002) *Tests and Assessments: Purposes and Quality.* Paper prepared for the first seminar of the Royal Society 14-19 science assessment enquiry.

BLACK, P. & WILIAM, D. (2003) In praise of educational research: formative assessment. *British Educational Research Journal* 29.5. 623-638

BLACK, P. (2004) *Issues in Assessment by Teachers*. Paper prepared for the Nuffield/ARG seminar on Assessment Systems for the Future, Cambridge, 12-13 January 2004

BOWE, R. & WHITTY, G. (1984) 'Teachers, Boards and Standards: The Attack on School-Based Assessment in English Public Examinations at 16+' in BROADFOOT, P. (ed) *Selection, Certification and Control*. London: Falmer

BRANTHWAITE, A., TRUEMAN, M. & BERRISFORD, T. (1981) Unreliability of Marking: further evidence and a possible explanation. *Education Review*. 33.1. 42-46

BROADFOOT, P. (ed) (1984) *Selection, Certification and Control*. Lewes: Falmer Press

BROADFOOT, P. (1994) 'Approaches to quality assurance and control in six countries' in HARLEN, W. (ed) *Enhancing quality in assessment.* London: Paul Chapman

BROADFOOT, P.M. (1996) *Education, assessment and society*. Buckingham: Open University Press

BROADFOOT, P., GOULDEN, D., LINES, D. & WOLF, A. (1995) *Evaluation of the use of set assignments in GNVQs: a report to the Employment Department.* University of London: Institute of Education/University of Bristol School of Education

BROADFOOT, P., JAMES, M., MCMEEKING, S., NUTTALL, D. AND STIERER, B. (1988) *Records of Achievement: Report of the National Evaluation of Pilot Schemes.* London: HMSO

BROWN, T. & BALL, S. (1992) *A report on the VCE verification process.* Melbourne: Victorian Curriculum and Assessment Board

BUCHAN, A.S. (1993) Policy into practice: internal assessment at 16+: standardization and moderation procedures. *Educational Research* 35.2 171-179

BURSTALL, C. (1994) 'Combining External and School-Based Assessments in England and Wales' in MAURITIUS EXAMINATIONS INSTITUTE (ed) *School-Based and External Assessments.* Papers from the 1993 conference of the International Association for Educational Assessment, Mauritius

BUTLER, J. (1995) Teachers judging standards in senior science subjects: Fifteen years of the Queensland experiment. *Studies in Science Education, 26*, 135-157.

CAPEY, J. (1995) *GNVQ Assessment Review*. London: NCVQ

CARTER, R.S. (1952) How Invalid are Marks Assigned by Teachers? *Journal of Educational Psychology* 43. 218-228

CHRISTIE, T. & FORREST, G.M. (1981) *Defining Public Examination Standards*. London: Macmillan

CHRISTOPHER, R., ROOKE, H.M. & HEWITT, E.A. (1970) *An experimental scheme of school assessment in Ordinary Level English Language: third report*. Occasional Publication 31. Manchester: Joint Matriculation Board

COHEN, L. (1974) The Stability of the Results of Agreement Trials. *Report to the Schools Council*

COHEN, L. & DEALE, R. (1977) *The Assessment of Teachers in Examinations at 16+* (Schools Council Examinations Bulletin 37) London: Evans/Methuen

CRESSWELL, M. (1987a) Describing examination performance: grade criteria in public examinations *Educational Studies* 13.3 247-265

CRESSWELL, M.J. (1987b) A more generally useful measure of the weights of examination components. *British Journal of Mathematical and Statistical Psychology*. 40. 61-79

CRESSWELL, M.J. (1996) 'Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches' in GOLDSTEN, H. & LEWIS, T. (eds) *Assessment: problems, developments and statistical issues.* London: Wiley

CUMMING, J. (1997) 'Defining contextualised assessment: authenticity, anchoring, simulation, abstraction, representation, camouflage' in AJAR, D. (ed) *New Horizons in Learning Assessment: Proceedings of the 1995 conference of the International Association for Educational Assessment.* Montreal: University of Montreal

DARLING-HAMMOND, L. (1994) Performance-Based Assessment and Educational Equity. *Harvard Educational Review* 64.1 5-30

DAUGHERTY, R. (1995) *National Curriculum Assessment: a review of policy* 1987-1994. London: Falmer

DEARING, R. (1995) *Review of 16-19 Qualifications, Interim Report: The Issues for Consideration.* London: SCAA

DELAP, M.R. (1995) Teachers' estimates of candidates' performances in public examinations. *Assessment in Education* 2.1 75-92

DfES (2003) *Excellence and Enjoyment: a Strategy for Primary Schools* www.dfes.gov.uk/primarydocument

DUGGAN, S & GOTT, G.(1996) Scientific evidence: the new emphasis in the practical science curriculum in England and Wales. *The Curriculum Journal* 7.1.17-33

DWECK, C. (1989) 'Motivation' in LESGOLD, A. & GLASER, R. *Foundations for a Psychology of Education.* Hillsdale, NJ: Earlbaum

ECCLESTONE, K. & HALL, I. (1999) Quality Assurance and Quality Control in Internal Assessment across Qualifications. Report to QCA from Dept. of Education, University of Newcastle

ECCLESTONE, K. & PRYOR, J. (2003) 'Learning Careers' or 'Assessment Careers'? The Impact of Assessment Systems on Learning. *British Educational Research Journal* 29. 4. 471-488

ELLEY, W.B. & LIVINGSTONE, I.D. (1972) *External Examinations and Internal Assessments.* New Zealand Council for Educational Research

EMERY, H., WILMUT, J. & MURPHY, R. (1998a) *Consistency in teacher assessment and the impact of SCAA guidance materials at key stage 2 in the non-core subjects*. Report to the Qualifications and Curriculum Authority.

EMERY, H., WILMUT, J. & FOX, R. (1998b) *Monitoring assessment at key stage 2.* Report to the Qualifications and Curriculum Authority.

ERAUT, M., STEADMAN, S., TRILL, J. & PORKES, J. (1996) *The assessment of NVQs* Research Report 4. Brighton: University of Sussex Institute of Education

FILER, A. (1993) Contexts of assessment in a primary school. *British Educational Research Journal* 19.1 95-107

FILER, A. (1994) 'Teacher assessment: a sociological perspective' in HUTCHINSON, D. & SCHAGEN, I. (eds) *How reliable is national curriculum assessment?* Slough: National Foundation for Educational Research

FIRESTONE,W.A. (1998) A Tale of Two Tests: tensions in assessment policy. *Assessment in Education.* 5.2. 175-191

GIPPS, C. (1994) 'Quality in Teacher Assessment' in HARLEN,W. (ed) *Enhancing Quality in Assessment.* London: Paul Chapman

GIPPS, C. (1994) *Beyond Testing: Towards a theory of educational assessment*. Lewes: Falmer

GIPPS, C.V. (1995) 'Reliability, validity and manageability in large-scale performance assessment' in TORRANCE, H. (ed) *Evaluating authentic assessment.* Buckingham: Open University Press

GIPPS, C. & CLARKE, S. (1998a) *Monitoring Consistency in Teacher Assessment and the Impact of SCAA's Guidance Materials at key stages 1, 2 and 3.* London: Qualifications and Curriculum Authority

GIPPS, C., CLARKE, S. & McCALLUM, B. (1998b) The Role of Teachers in National Assessment in England. Paper given at the AERA Conference, 1998

GIPPS, C., McCALLUM, B. and HARGREAVES, E. (2000) *What makes a Good Primary School Teacher? Expert Classroom Strategies*. London: Falmer

GOFF, P. & LEIMANIS, A. (1995) *Awarding Body Verification Procedures 1994-1995*. London: National Council for Vocational Qualifications

GOTT, R. (1986) The assessment of practical investigations in science. *School Science Review* 68. 411-421

GOOD, F.J. (1988) A method of moderation of school-based assessments: some statistical considerations. *The Statistician.* 37. 33-49

HACKER, R.G. & ROWE,M.J. (1998) A longitudinal study of the effects of implementing a National Curriculum on classroom processes. *The Curriculum Journal* 9.1 93-103

HAGER, P. & GONCZI, A. (1993) Attributes and Competence *Australian and New Zealand Journal of Vocational Education Research.* 1(1). 36-45

HALL, K., WEBBER, B., VARLEY, S., YOUNG, V. & DORMAN, P. (1997) A Study of Teacher Assessment at key stage 1. *Cambridge Journal of Education.* 27.1. 107-123

HARGREAVES, A. (1982) The Rhetoric of School-Centred Innovation *Journal of Curriculum Studies* 14.3 251-266

HARLEN, W., GIPPS, C., BROADFOOT, P. & NUTTALL, D. (1993) Assessment and the improvement of education. *The Curriculum Journal* 3.3. 215-230

HARLEN, W. & JAMES, M. (1997) Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education* 4.3 365-380

HARLEN, W. & DEAKIN-CRICK, R. (2003) Testing and Motivation for Learning *Assessment in Education.* 10.2. 169-208

HARLEN, W. (2004) *A systematic review of research evidence of the reliability and validity of assessment by teachers used for summative purposes.* An EPPI review conducted by the Assessment and Learning Research Synthesis Group. London: Institute of Education: EPPI Centre

HELSBY, G., KNIGHT, P. & SAUNDERS, M. (1998) Preparing students for the new work order: the case of Advanced General National Vocational Qualifications. *British Educational Research Journal* 24.1. 63-78

HEWITT, E.A. (1967) *The reliability of GCE O level examinations in English Language*. JMB Occasional Publication 27. Manchester: Joint Matriculation Board

HILL, P.W., BROWN, T. & MASTERS, G.N. (1993) *Fair and authentic school assessment: advice to the Board of Studies on verification, scaling and reporting results within the VCE.* Melbourne: Victoria Board of Studies

HILL, P.W., BROWN, T., ROWE, K.J. & TURNER, R. (1997) Establishing comparability of Year 12 school-based assessments. *Australian Journal of Education.* 41.1 27-47

HONG KONG EXAMINATIONS AUTHORITY (1998) *Review of the Public Examination System in Hong Kong*. Final Report

HOSTE, R. (1982) The construct validity of some Certificate of Secondary Education Biology examinations: the evidence from factor analysis *British Educational Research Journal* 8. 31-42

HOSTE, R. & BLOOMFIELD, B. (1975) *Continuous Assessment in the CSE.* Schools Council Examinations Bulletin 31.

HUSBANDS, C.T. (1976) Ideological bias in the marking of examinations. *Research in Education.* 15.17-38

JAMES, M. & CONNER, C. (1993) Are reliability and validity achievable in National Curriculum Assessment? Some observations on moderation in key stage 1 in 1992. *The Curriculum Journal* 4.1. 5-19

JOHNSON, C., WOLF, A. & BARTRAM, D. (1995) *External Assessment for NVQs* Discussion paper prepared for the Beaumont Review of NVQs

JOHNSON, D. (1998) Teacher Assessments and Literacy Profiles of Primary School Children in South Africa *Assessment in Education* 5.3. 381-412

KELLAGHAN, T., MADAUS, G.F. & AIRASIAN, P. (1982*) The effects of standardised testing*. Boston: Kluwer-Nijoff Publishing

KELLAGHAN, T. & GREANEY, V. (1992) *Using examinations to improve education: a study in fourteen African countries.* World Bank Technical Paper 165

KELLY, A. (1988) Gender differences in teacher-pupil interactions: a meta-analytic review. *Research in Education* 39. 1-23

KEOHANE, K.W. (1979) Proposals of a Certificate of Extended Education. London, HMSO

KLENOWSKI, V. (2003) *Rethinking assessment in higher education.* Paper given to the 29th IAEA conference, Manchester, October 2003

KORETZ, D. (1998) Large-scale Portfolio Assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education* 5.3. 309-334

LAVE, J. & WEGNER, E. (1991) *Situated learning: legitimate peripheral participation.* Cambridge: Cambridge University Press

LEARNING AND TEACHING FOR SCOTLAND (2004) http://www.ltscotland.org.uk/assess/

LEPPER, M.R. & HODELL, M. (1989) 'Intrinsic motivation in the classroom' in AMES,C. & AMES, R. *Research on Motivation in Education, Volume 3*. San Diego: Academic Press

LOCK, R. & FERRIMAN, B. (1989) OCEA – The development of a graded assessment scheme in science. Part IV The Pilot Phase. *School Science Review* 70(252). 103-112

MADAUS, G.F. & KELLAGHAN, T. (1993) The British experience with 'authentic' testing. *Phi Delta Kappan*, February 1993. 458-469

MASTERS, G. N. (1986). 'Comparing school based assessments: a hierarchy of procedures'. In *Theory, structure and action in education* Annual Conference of the Australian Association for Research in Education., University of Melbourne: Australian Association for Research in Education.

MASTERS, G.N. & McBRYDE, B. (1993) *An investigation of the comparability of teachers' assessments of student folios*. Brisbane: Tertiary Entrance Procedures Authority

MAXWELL, G.S. (1994) *School-based assessment in Queensland.* Brisbane: University of Queensland Graduate School of Education

MAXWELL, G.S. (1997) 'Teacher judgement of achievement standards in performance assessments'. in AJAR, D. (ed) *New Horizons in Learning Assessment*. Proceedings of the 1995 conference of the International Association for Educational Assessment, Montreal

MAXWELL, G. & CUMMING, J. (1998) *Reforming the culture of assessment: changes in teachers' assessment beliefs and practices under a school-based regime.* Paper presented to the conference of the International Association for Educational Assessment, Barbados.

McCALLUM, B., GIPPS, C., McALISTER, S. & BROWN, M. (1995) 'National Curriculum Assessment: emerging models of teacher assessment in the classroom'. in TORRANCE, H. (ed) *Evaluating Authentic Assessment.* Buckingham: Open University Press

MITCHELL, L. & BARTRAM, D. (1994) *The Place of Knowledge and Understanding in the Development of National Vocational Qualifications and Scottish Vocational Qualifications.* Competence & Assessment Briefing Series 10. Sheffield: Employment Department

MOON, J. (2000) *PDP Reflection in Higher Education Learning.* Working paper 4. York: LTSN Generic Centre http://www.ltsn.ac.uk/

MORGAN, C. (1996) The teacher as examiner: the case of mathematics coursework. *Assessment in Education* 3.3. 353-375

MORI/CDELL (2002) *Public examinations: views on maintaining standards over time*. Summary on QCA website http://www.qca.org.uk/products/95_1428.html

MOSS, P.A. (1992) Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research.* 62.3 229-258

MURPHY, P. (1995) Sources of inequity: understanding students' responses to assessment. *Assessment in Education* 2.3 249-270

MURPHY, J. (1974) Teacher expectations and working-class under-achievement. *British Journal of Sociology* 25. 326-344

MURPHY, R.J.L. (1978) Sex differences in examination performance: do these reflect differences in ability or sex-role stereotypes? *Educational Review* 30. 259-63

MURPHY, R.J.L. (1979) Teachers' Assessments and GCE Results Compared Educational Research 22.1 54-59

MURPHY, R.J.L. (1981) 'Statistical Moderation – A Critique' in ASSOCIATED EXAMINING BOARD *Combining teacher assessment with examining board assessment*. Report of a seminar. Aldershot:AEB

MURPHY, R.J.L. (1982) Sex differences in objective test performance. *British Journal of Educational Psychology* 52.213-19

MURPHY, R. & TORRANCE, H. (1988) *The Changing Face of Educational Assessment*. Buckingham: Open University Press

NEWTON, P.E. (2003a) The defensibility of national curriculum assessment in England. *Research Papers in Education*. 18.2.101-127

NEWTON, P.E. (2003b) Evidence-based policy making. *Research Papers in Education*. 18.2.137-140

NEWTON, P.E. (2004, personal communication) Teacher assessment for summative purposes: a Powerpoint presentation.

NITKO, A.J. (1995) Curriculum-based Continuous Assessment: a framework for concepts, procedures and policy. *Assessment in Education.* 2.3 321-338

NUTTALL, D. (1981) 'Criteria for successful combination of teacher assessed and external elements' in ASSOCIATED EXAMINING BOARD *Combining teacher assessment with examining board assessment*. Report of a seminar. Aldershot: AEB

NUTTALL, D. (1984) 'Doomsday or a New Dawn? The Prospects for a Common System of Examining at 16+' in BROADFOOT, P. (ed) *Selection, Certification and Control*. London: Falmer

NUTTALL, D. (1987) The Validity of Assessments. *European Journal of Psychology of Education* II.2 109-118

NUTTALL, D. & ARMITAGE, P. (1983) *The Moderating Instrument Research Project: a summary report*. Milton Keynes: The Open University School of Education

PAECHTER, C. (1995) 'Doing the Best for Students': dilemmas and decisions in carrying out statutory assessment tasks. *Assessment in Education* 2.1. 39-52

PATRICK, H. (2003) *Synoptic assessment: report for QCA.* Paper prepared for the Qualifications and Curriculum Authority by the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate

PEDRESCHI, T., CONNOR, J., THACKRAY, D. & WOLSTENCROFT, E. (1994) *Effective practice in assessment against the management standards.* Learning Methods Branch Research and Development Series; Report 25. Sheffield: Employment Department

PEDULLA, J.J., AIRASIAN, P. & MADAUS, G.F. (1980) Do teacher ratings and standardised tests of students yield the same information? *American Education Research Journal* 17. 303-307

PETCH, J.A. (1964) *School estimates and examination results compared*. Manchester: Joint Matriculation Board

PRYOR, J & TORRANCE, H (1996) Teacher-pupil interaction in formative assessment: assessing the work or protecting the child? *The Curriculum Journal* 7.2 205-226

QCA (1998) *Statistical Monitoring and Risk Assessment as Aids to Quality Assurance in NVQs.* Discussion paper

QCA (2003) *Response to the working group of 14-19 reform* http://www.qca.org.uk/ages14-19/2593.html

RADNOR, H.A. (1994) The problems of facilitating qualitative formative assessment in pupils. *British Journal of Educational Psychology*. 64. 145-160

RADNOR, H. & SHAW, K. (1995) 'Developing a collaborative approach to moderation' in TORRANCE, H. (ed) *Evaluating authentic assessment.* Buckingham: Open University Press

RAGGATT, P. & HEVEY, D. (1995) *Sufficiency of Evidence* Learning Methods Branch Research and Development Series; Report 32. Sheffield: Department for Education and Employment

ROBBINS, J. (1998) *Improving the dependability of examination coursework and assessment: a discussion paper.* Paper submitted to the Qualifications and Curriculum Authority

ROGERS, T.J. (1974) 'Coursework and Continuous Assessment' in MACINTOSH, H.G. (ed) *Techniques and problems of assessment – A practical handbook for teachers.* London: Edward Arnold

SADLER, R. (1989) Formative assessment and the design of instructional systems. *Instructional Science.* 18. 119-44

SCHOOLS COUNCIL (1965) *The Certificate of Secondary Education: School-based examinations*. Examinations Bulletin 5. London: HMSO

SECONDARY EXAMINATIONS COUNCIL (1985) *Coursework Assessment in GCSE*. Working Paper 2. London: SEC

SECONDARY EXAMINATIONS COUNCIL (1988) *Managing GCSE coursework in schools and colleges.* Working Paper 6. London: SEC

SHAPLEY, K.S. & 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

SIMPSON, M. (1990) Why criterion-referenced assessment is unlikely to improve learning. *The Curriculum Journal* 1.2 171-183

SMITH, G.A. (1978) *JMB experience of the moderation of internal assessments*. Occasional Publication 38. Manchester: Joint Matriculation Board

SPEAR, M.G. (1989) The relationship between standard of work and mark awarded. *Educational Research.* 31.1.69-70

STECHER, B. (1998) The Local Benefits and Burdens of Large-scale Portfolio Assessment. *Assessment in Education* 5.3. 335-351

STOBART, G., ELLWOOD, J. & QUINLAN, M. (1992) Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal* 18.3. 261-276

STOBART, G. (2003) 'Using assessments to improve learning: intentions, feedback and motivation' in RICHARDSON, C, (ed) *Whither Assessment?* London: Qualifications and Curriculum Authority

STRACHAN, J. (1997) 'Moderation of assessments in the national qualifications framework' in *New Horizons in Learning Assessment: Proceedings of the 1995 conference of the International Association for Educational Assessment*. Montreal: University of Montreal

STRAUGHAN, R. & WRIGLEY, J. (1980) *Values and Evaluation in Education*. London: Harper & Row.

TAYLOR, M. (1992) *The reliability of judgements made by coursework assessors*. AEB Research Report RAC 577.

TAYLOR, L. & LEE, S. (1994) The school assessment portfolio – has it a future? *British Journal of Curriculum and Assessment* 5.1 8-11

THOMAS, S., MADAUS, G.F., RACZEK, A.E. & SMEES, R. (1998) Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment. *Assessment in Education.* 5.2. 213-246

TOLLEY, H., GREATBATCH, D., BOLTON, J. & WARMINGTON, P. (2003) Improving Occupational Learning: The validity and transferability of NVQs in the workplace. DfES Research Report 425

TORRANCE, H. (1995) 'Teacher involvement in new approaches to assessment' in TORRANCE, H, (ed) *Evaluating Authentic Assessment.* Buckingham: Open University Press

TRAVERS, E.J. & ALLEN, R (1994) *Random sampling of student folios: a pilot study*. Brisbane: Board of Secondary School Studies

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE (1976) *School Examinations and their Function* Cambridge: UCLES

UNIVERSITY OF NOTTINGHAM (1995) *The reliability of assessment of NVQs.* Report to the National Council for Vocational Qualifications

WADDELL, J. (1978) *School Examinations*. London: HMSO

WALKER, D.A. (1979) 'The Standardisation of School Assessment' in SCOTTISH EDUCATION DEPARTMENT *Issues in Educational Assessment*. Edinburgh: Scottish Education Department Occasional Papers

WARD, M. (1982) *The Assessment by Teachers of Their Own Pupils for Public Examinations*. Research Paper RAC209. Aldershot: Associated Examining Board

WARMINGTON, P. & WILMUT, J. (2001) *The Roles of NVQ Internal Verifiers*. A report to the Department for Education and Employment from the Centre for Developing and Evaluating Lifelong Learning, University of Nottingham

WATSON, J.R., SWAIN, J.R.L. & McROBBIE, C. (2003) Australian Students' Discussions in Practical Scientific Inquiries. *International Journal of Science Education* (in press)

WEEDEN, P. & WINTER, J. (1999) *Learners' Expectations of Assessment for Learning Nationally*. Report to QCA. University of Bristol Graduate School of Education CLIO Centre for Assessment Studies

WILIAM, D. (1993) Validity, dependability and reliability in National Curriculum Assessment. *The Curriculum Journal.* 4.3 335-350

WILIAM, D. & BLACK, P. (1996) Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal* 22.5 537-548

WILIAM, D. (2003) National curriculum assessment: how to make it better. *Research Papers in Education*. 18.2. 129-136

WILLMOTT, A.S. & NUTTALL, D.L. (1975) *The Reliability of Examinations at 16+*. London: Macmillan Education

WILMUT, J. (1977) *The procedures available for the conduct and moderation of teacher assessment components in large entry examinations*. Unpublished research paper. Aldershot: Associated Examining Board

WILMUT, J. (1994a) *Assessment in NVQs in the Care Sector*. London: Joint Awarding Bodies

WILMUT, J. (1994b) *Quality assurance in NVQs in Business Administration.* Report to the National Council for Vocational Qualifications

WILMUT, J. (1995) *Assessment in NVQs in Child Care and Education*. London: Joint Awarding Bodies

WILMUT, J. (1997) 'Agreement Trialling for Professional Development in Assessment'. in AJAR, D. (ed) *New Horizons in Learning Assessment*. Proceedings of the 1995 conference of the International Association for Educational Assessment, Montreal

WILMUT, J., WOOD, R, & MURPHY, R. (1996) *A Review of Research into the Reliability of Examinations*. Discussion paper for the School Curriculum and Assessment Authority

WILMUT, J. (1999) *The use of internal assessment in qualifications*. A review of research for the Qualifications and Curriculum Authority

WILMUT, J. & MURPHY, R. (2001) *Securing Quality in Assessment: The Roles of Regulators, Awarding Bodies and Users.* Paper given of the 27th Annual Conference of the International Association for Educational Assessment, Rio de Janeiro, May 2001

WILMUT, J. & MACINTOSH, H. (2001) *Improving Reliability in Qualifications: Quality Assurance Procedures.* Report to the Qualifications and Curriculum Authority.

WOLF, A., BURGESS, R., STOTT, H. & VEASEY, J. (1994) *GNVQ assessment review project: final report*. Learning Methods Branch Technical Report 23. Sheffield: Employment Department

WOLF, A. (1995) *Competence-based assessment*. Buckingham: Open University Press

WOLF, A. (1998) Portfolio assessment as national policy: the National Council for Vocational Qualifications and its quest for a pedagogical revolution *Assessment in Education* 5.3 413-446

WOLF, A. (1999) 'Outcomes, Competencies and Trainee-centred Learning: the Gap between Rhetoric and Reality' in MURPHY, P. (ed) *Learners, Learning and Assessment* Chapman/ Open University

WOOD, R. (1972) *On Moderation* Paper presented to the GCS Secretaries Annual Conference, Royal Holloway College

WOOD, R. (1991) *Assessment and Testing*. Cambridge: University of Cambridge Local Examinations Syndicate

WOOD, R. & FERGUSON, C.M. (1975) Teacher assessment of practical skills in Advanced Level Chemistry *School Science Review* 57. 605-8

WOOD, R. & NAPTHALI, W.A. (1975) Assessment in the Classroom: What do Teachers Look For? *Educational Studies* 1.3. 152-161

WOOD, R. & POWER, C. (1987) Aspects of the competence-performance distinction: educational, psychological and measurement issues. *Journal of Curriculum Studies* 19. 409-24

WOOD, R., JOHNSON. C, BLINKHORN, S. & HALL, J. (1989) *Boning, Blanching and Backtacking: Assessing Performance in the Workplace*. Research and Development Series 46. Sheffield: Training Agency

