DOCUMENT RESUME

ED 367 712 TM 021 245

AUTHOR Harlen, Wynne
TITLE Concepts of Quality in Student Assessment.
PUB DATE Apr 94
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accreditation (Institutions); Criteria; Data Collection; Data Interpretation; Definitions; *Educational Assessment; Educational Quality; Elementary Secondary Education; Equal Education; *Evaluation Methods; Foreign Countries; Inspection; Professional Development; *Quality Control; *Reliability; Scaling; Statistical Analysis; *Student Evaluation; Test Construction; Test Use; *Validity
IDENTIFIERS Authentic Assessment; *Moderation; Quality Assurance; United Kingdom

ABSTRACT

This paper gives an overview of the methods of moderation, or quality assurance and quality control, as they may be more widely known, that are used to enhance the quality of student assessment. The discussion is based on the educational systems of the United Kingdom but is applicable to assessment in other countries. Quality in assessment is seen as the provision of information of the highest validity and optimum reliability suited to a particular purpose and context. Moderation procedures fall broadly into those concerned with adjustment of the outcome of the assessment to improve fairness (quality control) and those concerned with arriving at fair assessments (quality assurance). Approaches to quality control are: (1) the use of reference or scaling tests for statistical moderation; (2) the inspection of samples by post (e.g., sending a sample of work to an outside organization for an award); (3) the inspection by visiting moderators; (4) external examination; (5) teacher-requested moderation; and (6) group and consensus moderation of internal assessment. Approaches to quality assurance, which are usually made before the assessment, include definition of criteria, providing examples, approving entities carrying out assessment, visits of moderators, and group moderation. The quality of teacher assessments can be enhanced through moderation procedures that support professional development. (Contains 18 references.) (SLD)

Concepts of Quality in Student Assessment

Paper given at the AERA 1994 Annual Meeting, New Orleans as part of the symposium given by members of BERA on **Enhancing Quality in Student Assessment**

by

Wynne Harlen

Director, Scottish Council for Research in Education, 15, St. John Street, Edinburgh, EH8 8JR, Scotland

CONCEPTS OF QUALITY IN STUDENT ASSESSMENT

Wynne Harlen, Scottish Council for Research in Education

Introduction

This paper gives an overview of methods of moderation, or quality assurance and quality control, as they may be more widely known, which are used to enhance the quality of student assessment. In discussing the prior issue of what is meant by *quality* in assessment, a case is made for the importance of assessment by teachers as opposed to external assessment in the form of examinations or tests set and marked by external testing and examination agencies.

The quality of all kinds of assessment must, we believe, be judged by the same criteria and based upon evidence rather than assumption and tradition. Yet the UK is not alone in operating on the assumption that externally set tests are necessarily more dependable than internal assessment carried out by teachers. Recently there has been in England and Wales a quite explicit downgrading of assessment made by teachers, which is in sharp contrast to the confidence in teachers' assessment, shown, for example, in Germany (See Broadfoot, 1994, and in this symposium). Thus low reliability is not an inherent failing of teacher-based assessment. At the same time we recognise that teachers' assessments are sometimes more unreliable than would be the case if more resources were used to research the reasons and support measures to improve procedures.

Although the problem motivating this paper and other in this symposium is based in the educational systems and experience of the UK, we believe that the arguments and evidence will inform debate about assessment in other countries. For wherever problems in assessment practice are faced, and particularly where changes are towards broadening beyond traditional forms of tests and examinations, to encompassing skills and knowledge application rather than knowledge recall and to criterion-referencing, then the issues discussed here will arise.

The concept of quality in assessment

Assessment takes place in a wide range of contexts in education and for many different purposes. Purposes relating to individual pupils include informing the next steps in teaching, summarising achievement at a certain time, selection, certification and guidance. These all effect a pupil's immediate or longer-term opportunities and call for as much fairness and accuracy in the assessment as is possible. Pupils are also assessed for other purposes, such as surveys of national achievement and for research, where the results will not have a direct effect on the pupils assessed but, nonetheless, a dependable result is required. Such statements, although ones we can only agree with,

beg questions about what we mean by 'accuracy' and 'dependability' in this context and to what degree we can expect to achieve these qualities in educational assessment.

A comprehensive definition of assessment includes the processes of gathering, interpreting, recording and use of information about a pupil's response to an educational task (Harlen et al, 1992 p 217). A vast range of ways of assessing can be identified by combining different means of getting information (eg observing actions, listening, reading written work, studying products such as drawings and artefacts) with various kinds of task (eg written test or examination papers, practical tasks set externally or by the teacher, projects, tasks undertaken as part of normal class work). In the case of formal assessments there are further variations introduced by whether the marking is carried out by an external agency or by the teacher and the reference base used for interpretation (norm-referenced, criterion-referenced or pupil-referenced).

The reason for choosing one rather than another of this plethora of possible ways of assessing relates to the requirement for optimum *dependability*. This word needs to be understood in terms of the two interconnected concepts of reliability and validity since its meaning is otherwise ambiguous. The problem can be illustrated by asking the question: is a written science test more or less dependable than a practical science test? If dependable is taken to mean that the result would be similar if the test were repeated, then the written test would come out best. But if it means the one that is the most dependable measure of how practical science tasks are tackled, then the reverse would be the case. So these two aspects have to be disentangled.

The concept of *reliability* of the result of an assessment refers to the extent to which a similar result would be obtained if the assessment were to be repeated. The aspect of an assessment which refers to how well the result really reflects the skill, knowledge, attitude or other quality it was intended to assess is described as its *validity*. These are two distinct attributes of an assessment, as the example of the science test suggests, but they are also interconnected. An assessment which is low in reliability, that is, gives widely varying results if repeated, can hardly have a high validity, since it will be unclear just what is being assessed. Conversely it is difficult in some cases to have high reliability *and* high validity, since the requirements of high reliability lead to close specification of task, response mode, means of gathering information and interpretation and these are often incompatible with high validity. For example, a standardised reading test has high reliability when children are required to read isolated words, free of context. The results, however, are a less valid assessment of how well children can read when using context to help them, an ability which would be tested by a more realistic (but less reliable) reading task.

The highly significant, and sometimes overlooked, consequence of this argument that validity and reliability can never both be 100%, is that we must recognise

assessment is never 'accurate' in the way that the word is used in the context of measurement in the physical world. Assessment in education is inherently inexact and it should be treated as such. We should not expect to be able to measure pupils' abilities with the same confidence as we can measure their heights. This in no way makes educational assessment useless. It means that the interpretation of assessment results should be in terms of being an indication of what pupils can do but not an exact specification. At the same time this is no excuse for failing to make every effort to ensure that both validity and reliability are at an optimum level, for information which is low in these qualities cannot be used effectively and, if used, could lead to injustice.

Both reliability and validity have to be considered in relation to the *contexts* and *purposes* of assessment. A highly reliable assessment but one which is time consuming or demanding of resources will be of little use to a teacher who wants information about pupils on a regular basis with minimum interruption of normal work. In such circumstances *quality* in assessment means an assessment made and interpreted on the spot which provides the type of information required (high validity) and with the greatest degree of reliability possible in the circumstances. The intended use of the information in this case means that reliability is not the foremost consideration. However, had the purpose been to provide an assessment of course work as a contribution to an external award, the burden on reliability could be greater. In both cases, however, the value depends on the ability of the teacher to gather and interpret the information with the required rigour and respect for evidence. Good assessment thus depends on the use and development of these skills. These sorts of consideration lead to the proposition that *quality in assessment is the provision of information of the highest validity and optimum reliability suited to a particular purpose and context.*

Why assessment by teachers?

The usefulness of an assessment is directly related to its validity, providing it is not so low in reliability as to call this into question. This is saying no more than that if we want information about, say, a student's ability to solve mathematical problems, we need to have assessed them trying to solve mathematical problems. Results of a reliably marked arithmetic test will not be useful for this purpose. Thus as a priority for ensuring quality in assessment, validity has first consideration. Thereafter every effort has to be made to increase reliability and in this pursuit tightness of specification of task and interpretation adopted is the highest that is still compatible with required validity.

It is not difficult to see that, for the purpose of assessing many attributes, formal tests and examinations do not compare very favourably with less restricted methods of assessment. This is well demonstrated by the list of skills and abilities identified by the Secondary Examinations Council (SEC) as needing to be assessed

through course work in the General Certificate of Secondary Education (GCSE), normally taken by 16 year-olds:

a) the ability to use and develop techniques for making and recording accurate observations, in the context of, for example, fieldwork or experimental work;

b) research skills, including the ability to organise the systematic collection and ordering of pertinent information; familiarity with and us of a wide range of sources; the ability to distinguish sources of different status in weighing evidence – for example, primary and secondary sources;

c) interactive skills (responding appropriately to the consequences of an earlier action); such interaction may be with people, information sources (including information technology), tools or concrete materials;

d) the ability to find a role and co-operate with others in an activity;

e) motor skills including manipulation of apparatus, operation of machinery, and marking out and processing of materials;

f) skills involving a sense of timing; the ability to 'think on one's feet';

g) the exercise of safety awareness;

h) the ability to design, conduct and evaluate a simple experiment or survey to test some hypothesis or illuminate some issue;

i) the ability to make a simple theoretical model of a 'real-life' situation and to test and refine the model by examining both it and the real-life situation further;

j) the determination and ability to sustain a chosen study from conception to realisation;

k) attainment in tasks which, by their nature, require time for exploration; investigational, planning and design activities where several approaches may need to be considered before a specific solution is developed; activities where several resource constraints (such as those of cost, time and skill) have to be investigated and weighed before a solution is pursued;

l) attainment in areas where it is desirable to allow time for reflection, for example, in articulating a thoughtful personal response to the expressive arts or to religious experience or in teaching an objective and informed view of some current social or moral issue;

m) skills of adaptation and improvisation in the widest sense: the ability to restructure information or modify objects to suit immediate needs; the ability periodically to review the progress of a long-term enterprise (such as a scientific experiment, a piece of planning or a craft or agricultural project) and to change tactics if necessary; the exercise of awareness of possible sources of difficulty or error.

Other items could be added to the list which are equally important objectives of education and ones most valued in the modern world, as indicated by their inclusion in concepts such as 'enterprise', higher order thinking abilities and transferable skills. Unless the assessment includes these attributes they will be under-valued and under-developed.

The key feature relating to the validity of assessment of any kind, but which is particularly relevant to our present argument, is that of *opportunity* for the students to show that they have the abilities, skills, etc in question. So, for each of the items in the above list, we should ask 'what conditions/situations give students opportunity to show the ability or skill?' It is not difficult to realise that most require assessment to be made over an extended time and in conditions which approach 'realistic' situations; many cannot validly be assessed in short periods of time under examination conditions and none can be validly assessed by written items alone. Since situations which provide opportunity for these abilities to be used and developed must exist in teaching, then it follows that course work can also provide opportunity for valid assessment.

Similar points apply to the assessment of young children, where the notion of opportunity is equally useful. At the simplest level, a child may be able to draw and discuss his or her ideas with the teacher but not to write them down. Thus a written task would not provide opportunity for the ideas to be assessed whilst the normal classroom activities would do so. The point is wider, however, than just avoiding the use of skills (such as reading and writing) which are not under test. It extends to the meaning children perceive for a task, their past experience of it, their interest in it. It is well established that these things influence performance. Thus giving the same task to children under the same conditions is not necessarily providing equal opportunity for them to show what they can do or what they know. A more valid assessment would be made across a range of situations, such as can be done by a teacher assessing as an on-going part of teaching.

However, the matter of reliability must be faced, for, as stated earlier, an unreliable assessment is not only of little use but can be unjust. The endeavour to increase reliability is common to all methods of assessment but the context and purpose of assessment will affect the degree of priority given to reliability. A higher priority is necessarily accorded to it when the measurement of attainment contributes to the certification of the student or, in aggregated form, to an evaluation of the performance of teachers and schools. Where a teacher is assessing his/her pupils in order to feed back into helping their learning, reliability need not be a major consideration.

It is recognised that constant attention has to be given to the reliability of external examinations with papers being remarked and results adjusted. In these circumstances, the tasks to which students respond are pre-specified and the conditions under which students respond are controlled. Procedures are required both to monitor consistency of presentation of the tasks in practice and to standardise marking but the inherent uniformity of tasks and conditions in examinations suggests dependability. However, here is considerable evidence that the higher reliability of examinations over teacher-based assessment should not be take for granted (see Satterly, 1994).

Nevertheless, traditionally there has been more confidence in external examinations in the UK and less in course work or teacher assessment. The key role of the latter, which we have argued in terms of validity and opportunity, indicates that it is important to increase confidence in this area. The means of achieving this is through various procedures which until recently were described as various forms of moderation. The varieties of moderation are about as extensive as the variety of methods of assessment and the rest of this chapter attempts to review the range and to suggest a framework for describing and comparing different approaches.

Moderation: quality assurance and quality control

Moderation procedures have been devised in order to reduce those sources of error which are seen to be greatest in particular circumstances whilst at the same time preserving validity of assessment as required for quality in assessment. The sources of error include variation in the demand or opportunity provided by the tasks undertaken by students, differences in interpretation of performance criteria or marking schemes and the intrusion of irrelevant contextual information in making judgements (see James, 1994, and in this symposium). Categorising the variety of measures taken to reduce such errors risks the oversimplification of any classification system, and the usual caveats apply here. Against these disadvantages have to be placed the advantage that categorisation provides a basis for comparing the pros and cons of various methods.

At an initial level of categorisation, moderation procedures fall fairly readily into one of two kinds:

(i) those concerned essentially with adjustment of the outcome of assessment to improve fairness for groups and individuals

(ii) those concerned with the process of arriving at fair assessments for groups and individuals, which will, in some cases, extend to opportunities to learn as well as to be assessed.

The first of these takes place after the assessment has been made and is designed to ensure fairness by adjusting results where there seems to be inconsistency or systematic

differences in the way procedures have been followed. For example, there may be a 'reference test' given to all students against which course work or teacher assessments are compared. The marking of the latter may be adjusted to put students from one teacher in the same rank order as given by the reference test. Moderation of this kind is also called into play when results of the two forms of assessment (teacher's assessment and standard task performance) in the National Curriculum Assessment in England and Wales are combined. According to the procedures used in 1991 and 1992, teachers could request moderation if they considered that accepting the standard task result would be unfair.

Types of moderation of the second kind take place *before* the assessment is completed. They are designed to improve the process of assessment in order to "ensure that consistency has been achieved, rather than to impose it on an otherwise inconsistent assessment system" (NISEAC, 1991, para 10.1). The quotation is from a document which proposed moderation procedures for the national assessment in Northern Ireland which illustrate well this formative approach. The kinds of action proposed involved teachers meeting to discuss pupils' work both within one school and with teachers from other schools. Visits of moderators to schools were also proposed so that any systematic variation between teachers would be spotted. The overall purpose was, however, not to adjust marks and settle disputes, but to improve the quality of the assessment process.

The distinction between (i) and (ii) can be recognised as the distinction between quality control and quality assurance (eg Wiliam 1992). In the industrial model, quality control is the process of weeding out the imperfect products, meaning those which fall outside certain tolerance limits. Quality assurance constantly monitors the steps in arriving at the product and, in making sure that all processes are optimally carried out, theoretically prevents imperfect products. This analogy might suggest that the distinction might be better described as concern with product or with process. Although useful, in the assessment context there is more interaction between impact on process and product. Not only is attention to improving the assessment process justified in terms of a more reliable product but the discussion of an possible change to an assessment outcome during the moderation process can have an impact on the process of arriving at future decisions. The role of moderators in the assessment of 7 year olds in the National Curriculum Assessment, described by James (1994, and in this symposium) illustrates this interaction and the dilemma for the moderator of having a dual role in relation to both teachers' assessments and standard task administration.

The distinction between a quality assurance procedure and a quality control procedures does not reside inherently in the nature of the procedure; the categorisation must be made in terms of the purpose and effects of the procedure. To take an example

well illustrated in later chapters, the meeting of teachers in groups to discuss students' work may have a quality control purpose, if the result is that the judgements made about the work discussed are being scrutinised, or it may have a quality assurance purpose if examples of students' work are discussed in order to clarify the meaning of criteria but has no role in the assessment of the particular examples. Thus group moderation appears under both quality assurance and quality control in the following accounts. Of course in some cases the procedures will have a dual function, since when teachers discuss specific cases it will almost inevitably have an impact on their own understanding of the criteria, but an essentially quality assurance process will not have a quality control function. Similarly almost all quality control procedures can have a quality assurance function if the results are fed back to pin-point the source of error and to help in removing it.

Moderation procedures of quality assurance or quality control are used to improve the essentially imperfect process of assessment, but themselves vary in efficiency and in other important features such as cost. Thus not all approaches to moderation are equally useful and viable and it is necessary to identify the features of quality in moderation, just as we have for quality in assessment. We shall consider this after first providing some brief accounts of the main approaches, taken under the headings of quality control and quality assurance.

Approaches to quality control in assessment

The common feature shared by procedures in this category is that they occur after the event. They vary, however, in other respects, which emerge from the following examples.

Use of reference or scaling tests for statistical moderation

This is a device for adjusting students' marks using results of an externally marked test taken by all students. It is used in some cases to adjust assessments made by teachers in order to compensate for systematic variations in teachers' judgements. The Australian Scholastic Aptitude Test (see Broadfoot, 1994, and in this symposium) provides an example; others are described in Newbould and Massey (1979). In these cases the rank order of students assessed by one teacher or school stays the same but all scores may be moved up or down. It is also used where comparisons have to be made between students who have been examined, either by internal or external assessment, in different subjects of which some may be easier than others, as in the calculation of the tertiary education rank in New South Wales. In this case the rank order of students is likely to be changed. In these and other variants on the procedure, the student teacher

and institution have no control of what happens. It happens automatically and without any participation of the teacher beyond supplying the raw scores.

Inspection of samples by post

Here work assessed internally by teachers for an external award is sampled by the examination centre, usually the awarding body, to check that tasks have been set as required and that they have been marked and graded according to instructions (see, for example, the account of the Joint Matriculation Board's procedures by Smith, 1978). Precise instructions for selecting samples are normally given, so the teacher does not exercise choice in the matter. Inspection of samples is the principal method currently used by examination boards for moderation of coursework assessment (see Black, 1994). The procedure applies where written or other products on paper are used in assessment; thus where it is the only form of moderation there is a tendency for the assessment to be restricted to such forms.

Inspection of samples by visiting moderators

In principle this is similar to inspection by post but in practice the face to face contact between teacher and moderator or verifier facilitates professional discussion with reference to processes as well as to products. A wider range of products can be included in the assessment and moderation, although since the visit takes place after the work has been produced it cannot include ephemeral products or processes of working. The visiting moderation procedures of the Scottish Examination Board and several other similar Boards illustrate this approach. The cost of such exercises is a considerable deterrent to their use except for a sample of institutions at any one time.

External examining

This is a further variation of moderation by inspection of samples. It is widespread practice in higher education, developed to prevent variation in standards of awards between institutions which grant their own degrees, diplomas, certificates, etc. Examinations in these institutions are internally set and internally marked and so are as much in need of moderation as are continuously assessed components which are clearly more dependent on the judgement of individual lecturers. Examiners sometimes comment on papers or tasks set and on procedures, but their chief function is to comment on the standard of work of the students who pass or are given various grades and in some cases this results in the adjustment of grades.

Teacher requested moderation (appeals)

It is a feature of almost every certification procedure that an appeal can be made when the outcome is not what was expected. The appeal is usually on the grounds that the examination has not been carried out correctly and, when marking is at fault, the result can be changed. However, up to 1993 the National Assessment arrangements allowed appeals on grounds other than that the assessment was faulty. Because the result of teachers' assessment and the result of the standard test or task had to be combined, with the latter taking precedence, there was the possibility that differences would arise due to the nature of the two types of information. One of the tasks of local education authority moderators was to consider such appeals, although the process was initiated by the teacher (see Daugherty, 1994).

Group/consensus moderation of internal assessment

This involves the review of work which has been internally examined either as part or as the whole of an examination. The focus is the extent of agreement with the grade or scale point assigned to particular pieces of work by teachers. As already mentioned the process is little different from group moderation for quality assurance purposes, but the intention here is to ensure that grades have been assigned as agreed rather than to affect the process of arriving at the grading in the first place. The moderation procedures adopted in Queensland include this type (see Broadfoot, 1994, and in this symposium).

Approaches to quality assurance in assessment

Turning now from procedures where the main purpose is quality to control to those where it is quality assurance, the common feature of procedures in this category is that they attempt to increase dependability of teachers' assessments. They usually take place before the assessment is made but can operate in a post-hoc fashion. In the latter case the quality assurance function is distinguished from one of quality control only by the use made of the information. For example, whilst reference tests are no longer used to adjust marks in public examination in the UK, they may be used to draw attention

> to those cases where the locally determined marks are different from what might be expected on the basis of the same candidates' performance on the national yardstick. Such cases can then be investigated in detail by visiting verifiers, who are in the best position to draw a distinction between two major possibilities: local interpretation of standards out of line with national ones, or a level of performance on the local component genuinely different from what was expected as a result of performance on the national yardstick (perhaps as a result of extra emphasis or reduced emphasis on the local component, in terms of time, interest or resources)
>
> Nuttall and Thomas, 1993, p6

As this passage indicates, quality assurance procedures have to be concerned with both validity and reliability and so the focus is on the opportunities for learning and assessment and on increasing shared understandings of assessment criteria and

procedures. There is a considerable variation, however, in the extent to which these concerns can be addressed in a single approach. Thus the quality assurance procedures used in several cases comprise a combination of approaches. For the purpose of exemplifying possible procedures, however, the approaches used are considered separately.

Defining criteria for assessment

The provision of criteria to be applied in all schools by the government, in the form of statements of learning outcomes in the national curricula of England and Wales, Scotland and Northern Ireland, and the provision of national criteria for vocational qualifications by the National Council for Vocational Qualifications (NCVQ) in England and Wales and its counterpart in Scotland (SCOTVEC), were intended to provide a basis for uniformity in assessments made in the school, college or work-place. The 'intrinsic moderation' which is brought about by close specification of examination syllabuses and marking schemes, as in South Australia (see Broadfoot, 1994, and in this symposium) is a further example.

However, the reliability of criterion-referenced, or competence-based, assessment depends upon the clarity and ease of application of the statements of performance. The familiar dilemma facing those who define criteria is that the more specific and unambiguous the statements are, the more numerous and the less meaningful in terms of the abilities and understandings which are the real educational aims. Conversely the more the statements reflect complex, and particularly higher level, learning outcomes the more difficult they are to use reliably in assessment. The approach to quality assurance of providing criteria is, therefore, often accompanied by a constant revision aimed at improving the specification of criteria, attempting to avoid the problems of being too general or too specific. Examples are changes made to the National Curriculum statements of attainment in response to the unmanageability of the curriculum for assessment purposes and the changes made to Scottish National Certificate performance criteria as part of a revision of quality assurance procedures (Scottish Vocational Education Council (SCOTVEC), 1991). The provision and revision of criteria, whilst usually involving representatives of those who have to use them, is not open to negotiation with teachers and other users. Thus it is somewhat remote from influence by teachers and equally, of itself, weak in influence on teachers' assessments of individual students.

Exemplification

The provision of examples of pupils' work which has been assessed, preferably with a commentary on particular features used in making the judgement, enables abstract

criteria to be made specific. Good examples also indicate the type of task which provides opportunity for pupils to develop and to make evident their achievement of skills and understanding. Examples include the School Examinations and Assessment Council's (SEAC's) publications *Children's Work Assessed (Key Stage1)* and *Pupils' Work Assessed (Key Stage 3)* and the proposed assessment handbook to accompany the New Zealand achievement initiative (Broadfoot). Such examples can be used in group discussions similar to agreement trials (see below) but can also be used by a teacher individually. Unlike agreement trials, however, exemplification does not deal with work carried out in a context familiar to teachers and in that respect may have less impact on teachers' reflection on their own practice.

Approval of institutions/centres

This approach to quality assurance is a process by which the body responsible for certain awards approves an institution or centre as one which can provide the course or training and can carry out the assessment related to these awards. Institutions or centres are visited, course or training documents are reviewed, qualifications of staff are vetted and resources are inspected. Assessment procedures are included, although there may be other moderation procedures required by the awarding body in relation to quality control. Examples of this approach are the visitations which the Council for National Academic Awards used to carry out in approving courses and institutions and the 'quality auditing' of centres introduced by SCOTVEC from 1992. The approval of a centre may have implications for the locus of responsibility for ensuring reliability of assessments, since in some cases this will be devolved to the institution or centre. It will also affect the assessment process directly or indirectly by ensuring some standardisation of procedures across institutions or centres.

Visits of verifiers or moderators

This refers to visits carried out to observe the way in which assessment is carried out rather than to discuss products after the event, although in certain circumstances the two may be combined. Examples include the visits of moderators to schools during the administration of standard tasks as part of national assessment (NCA) and visits of verifiers to work places or colleges where National Vocational Qualifications (NVQ) assessments are being made. The emphasis is upon the procedures being implemented and the way in which criteria are being applied. Whilst the latter involves consideration of particular pieces of work or performance, the purpose is to inform the interpretation of criteria statements rather than to arrive at an agreed assessment in the cases that happen to be discussed. In this connection, it has been reported that moderators rarely changed Key Stage 1 teachers' judgements during NCA visits in order to preserve

teachers' confidence (James and Conner, 1993). The intention is to influence the assessment process and by doing so ensure greater reliability of assessment of students other than those whose performance was considered in the visit.

Group moderation

Also known as agreement panels, these are meetings of teachers or lecturers at which examples of work are discussed, the purpose being to arrive at shared understandings of the criteria in operation. The in-service function is foremost and the benefit greatest when teachers feel able to express their judgements and justify them openly, so that different conceptions and assumptions can be addressed. Membership may be from one school or several and groups led by an appointed local authority moderator or by a teacher. Persuasive arguments can be made for both inter-school and intra-school panel meetings; indeed both were proposed by NISEAC as part of the Northern Ireland national assessment arrangements (NISEAC, 1991). Inter-school meetings have a greater impact on reliability at the system level than intra-school meetings but are more costly.

Quality in moderation

Quality in moderation procedures can be considered using the same concepts of validity and reliability as for quality in assessment. In effect this would be saying that the approach(es) used would be the most relevant in relation to sources of variability (ie the most valid) and would have consistent and repeatable effects on different occasions (reliability). Although this application of validity and reliability has been pursued effectively in relation to secondary school examinations (SSABSA 1988) it is perhaps rather too theoretical for our purposes here. (For example, very little information exists about the reliability of moderation procedures even when restricted to those used in such examinations.) Moreover moderation must be concerned with all parts of the assessment process, from planning to product, and with what is in the teacher's mind as well as with public procedures. Thus we argue that improvement in teacher assessment is intimately related to professional development and that therefore this should be a major factor in deciding how to go about moderation.

It is evident that professional development is time-consuming and costly and these aspects cannot be ignored. But statistical and bureaucratic procedures also have their cost and thus should also be judged in these terms. More generally, we can draw out from the discussion of moderation procedures a number of aspects which should be considered in comparing their advantages and disadvantages. Figure 1 does this in terms of six aspects: the extent to which the moderation procedures are bureaucratically controlled, their contribution to professional development of teachers, their demands in

terms of time and costs and their impact on the process and the product of assessment. Judgements, which must be regarded as rough guidelines only, are indicated by a rating from one to three. (The justification for these ratings derive from the evidence of procedures presented at greater length in Harlen (ed), 1994).

Figure 1 **Profiles of moderation procedures**

Procedure	Bureauro cratic	Contribut ion to profession al dev	Time	Cost	Impact on process	Impact on product
Statistical - reference tests	▪▪▪	-	•	•	-	▪▪▪
Inspection of samples	•	•	•	•	-	▪
External examining	•	•	▪	▪▪▪	•	▪
Group moderation of grades	•	▪	▪	▪▪▪	▪	▪
Defining criteria	-	▪	▪	•	•	▪
Exemplification	-	▪	•	•	▪	▪
Centre approval	-	▪	•	•	▪	▪
Moderator visits	-	▪	▪▪▪	▪▪▪	▪▪▪	•
Group moderation	-	▪▪▪	▪▪▪	▪▪▪	▪▪▪	▪

Looking across the rows of Figure 1 gives a set of profiles of moderation procedures. In each one, not surprisingly, costs and time tend to be similar. It is also relatively easy to pick out those we have described as quality control since these have a high impact on the product of assessment and small impact on the process. However, those described as quality assurance have an impact on both process and product, supporting the claim by Gipps (1994) that a focus on quality assurance - improving the quality of the process - will inevitably also lead to an improved quality of product and hence greater consistency in standards and confidence in assessment results.

Looking within columns provides a means of comparing approaches in terms of specific characteristics, although the interdependence of the characteristics must not be forgotten. As might be expected, this shows that procedures involving the movement of people are expensive. However both this and the time aspects are also dependent on the scale of the operation. For example, external examining in higher education may be acceptable in terms of time and cost, but visits of moderators to all schools for national curriculum assessment would not be because of the much greater number of schools than of higher education institutions.

As with assessment procedures, so with moderation approaches, their acceptability will depend upon tradition and social values as much as on rational argument. The differences, and the choice of one or another, can be described but not

explained. Thus the statement about quality of moderation preferred by educators in the UK, where teachers generally value participation in decision making and seek to use professional judgement in making assessments, would be: *the process which optimises the reliability of an assessment at a cost which is balanced by the benefits in terms of the purposes of the assessment and contribution to professional development* . This preference is reinforced by embracing objectives of education which extend far beyond the type of knowledge and skills development which can be assessed by formal written tests. Thus, if the professional development element is seen as a *sine qua non,* then Figure 1 suggests that group moderation would be judged as being of the highest quality, all other things being equal. However, the other aspects, particularly the practical aspects of time and cost, cannot be ignored.

The alternatives which have lower cost but fairly high impact on process and product and on professional development appear to be 'exemplification' and 'centre approval'. The value of exemplification has been noted in several studies (see, for example, Gipps and James in this symposium), whilst centre approval is seen as the appropriate course in further education in Scotland (Black, 1994). Making schools into 'approved centres' of assessment will require them to have in place procedures for ensuring quality assurance and control and the appointment of assessment coordinators has already taken place in some primary as well as secondary schools. The increase in professional responsibility for quality in assessment is to be welcomed. However, in the context of schools being increasingly placed in competition with one another, there is a danger that it will be seen necessary for public confidence to provide some measure of external control to ensure consistency of standards across schools. Unfortunately the means of dong this through school inspectors or visiting peer groups is not seen to be as effective as using statutory tests and it seems that a return to beaurocratic control of teacher assessment through national tests is the Government's preferred model.

In our view it is possible to enhance the quality of teachers' assessments through moderation procedures which support professional development. By doing so we would achieve assessment results which would give dependable information about pupils' and students' performance across the wide range of aims of education. To do otherwise sets up a self-fulfilling prophecy which lowers teachers' professional status and so reduces public confidence in their judgements.

References

Black, H. (1994) Competence and quality: some research on the quality of assessment in Further Education in Scotland. In W. Harlen (ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London

Broadfoot, P. (1994) Approaches to quality assurance and control in six countries. In W. Harlen (ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London

Daugherty, R (1994) Quality Assurance, teacher assessment and public examinations. In W. Harlen (ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London

Dearing, R, (1993) *The National Curriculum and its Assessment* Interim Report. NCC and SEAC, York and London.

Harlen, W, Gipps, C, Broadfoot, P and Nuttall, D (1992) Assessment and the improvement of education, *The Curriculum Journal,* Vol 3, No 3, 215-230

Gipps, C. V. (1994) Quality in Teacher Assessment. In W. Harlen (Ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London.

James, M (1994) Experience of Quality Assurance at Key Stage 1. In W. Harlen (ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London

James, M and Conner, C. (1993) Are reliability and validity achievable in National Curriculum Assessment? Some observations on moderation at key stage one. *The Curriculum Journal,* 4 (1)

Newbould, C.A. and Massey, A. J (1979) *Comparability using a Common Element* Test Development and Research Unit, Cambridge

NISEAC (1991) *Pupil Assessment in Northern Ireland* Advice to the Lord Belstead, Paymaster General NISEAC, January 1991

Nuttall, D.L and Thomas, S. (1993) *Monitoring Procedures Based on Centre Performance Variables.* Report No 11. A Technical Report Published by the Employment Department's Methods Strategy Unit, Sheffield.

Satterly, D. (1994) Quality in external assessment. In W. Harlen (ed) *Enhancing Quality in Assessment.* Paul Chapman Publishing Ltd, London

School Examinations and Assessment Council (SEAC) (1991) *Children's Work Assessed (KS1)* and *Pupils' Work Assessed (KS3),* London, SEAC

SCOTVEC (1991) *Quality Development Programme.* Policy Paper. SCOTVEC, Glasgow

SEC (1985) *Coursework Assessment in GCSE.* Working Paper 2. SEC, London

Smith, G. A. (1978) *The JMB Experience of the Moderation of Internal Assessments* Joint matriculation Board (Occasional paper 38), Manchester.

SSABSA (1988) *Information Booklet No. 2 Assessment and Moderation,* South Australia, SSABSA

Wiliam, D (1992) Some technical issues in assessment: a user's guide. *British Journal of Curriculum and Assessment,* 2 (3), 11-20