A REVIEW OF RESEARCH INTO THE RELIABILITY OF EXAMINATIONS

A discussion paper prepared for the School Curriculum and Assessment Authority by

John Wilmut
Robert Wood
Roger Murphy

May 1996

Contact address:
Professor Roger Murphy, School of Education, University of Nottingham, University Park, Nottingham NG7 2RD.
Tel 0115 951 4477; fax 0115 951 4475

Preamble

This paper has been written as a response to a specification from the School Curriculum and Assessment Authority, requesting a review of research into the reliability of examinations, as follows.

The work will involve a review of relevant research on reliability. including that on GCE, CSE and GCSE examinations, and other appropriate assessment carried out in the UK or elsewhere.

The review must consider both the reliability of overall examinations and that of individual components, including coursework and externally set examinations. This review will form the basis of a report to SCAA that will

i. consider the meanings of 'reliability';
ii. describe techniques by which reliability can be measured;
iii. survey relevant research;
iv. make recommendations on possible further work that SCAA could conduct on reliability.

In the course of preparing the paper we contacted research officers at the examinations boards in England, Wales and Northern Ireland, asking for information about recent work on examination reliability of which they were aware. As a result of this we were supplied with copies of a number of reports, some of which are yet to be published. We are most grateful for this help.

Introduction

Basis for the discussion of reliability

Our discussion of the reliability of examinations draws on a number of types of sources.

- There has been a considerable amount of relevant work in the general field of psychometrics, since public examinations use many of the same methods, though often operating on a larger scale. Much of this work uses classical definitions of reliability to which has been added the more recent work on generalisability.

- A considerable body of literature has developed around examinations themselves, particularly in the matters of marking and grading reliabilities, where both of these are strongly affected by the broad scope of most examinations (wide range of domains tested), the scale upon which they operate and the relatively short time between the taking of the examination and the publication of results. A good deal of this literature focuses on the characteristics of different types of written paper and on different approaches to marking.

- A much smaller body of literature has developed around the operation of teacher assessment, which is relevant to examinations since coursework has formed a significant part of British examinations for some years. However, there is an increasing literature on the work of teachers as assessors which may illuminate their work in an examinations context.

- Recent literature on national curriculum assessment highlights the need for some alterative or additional approaches to reliability, when this is to be applied to assessments made in relation to criterion statements. Moves towards the specification of examination syllabuses in the form of outcome statements, or the development of more explicit criteria for marking or grading, would immediately make this work relevant to examinations.

- There is a slowly expanding body of literature on the reliability of assessment in competence-based systems of assessment. In Britain this is largely referred to vocational qualifications, and may become relevant as a more coherent approach to post-16 assessment develops.

The present discussion paper

Our discussion paper is organised in four main sections following the remit from SCAA; these are

- The meanings of reliability

- Techniques by which reliability can be measured
- Research into examination reliability
- Recommendations for further work

Within these four sections we have incorporated discussions of related areas such as the reliability of national curriculum assessment and the reliability of competence-based assessment, judging that work in these areas is likely to have some impact on future work in examinations.

The marking bibliography

We have provided a selected list of references which support our main discussion. However, the literature on reliability is vast, and only part of it is directly relevant to examinations. One portion, which is of special interest, deals with marking and with aspects of between and within examiner variation, all of which we have dealt at length within the discussion paper. However, in order adequately to reflect the range of work which has been done in this area we have provided a bibliography on the reliability of marking, but have kept it as a separate document, since it would have overwhelmed the main report had the two been merged.

The meanings of reliability

Introduction

For the public at large, there are, perhaps only two senses in which examination reliability matters. The first is that the grades which candidates receive from a particular examination should be exactly the grades they deserve, not one grade higher nor one grade lower. Operationally, this means that, were the examination to be repeated, candidates would emerge with the same grades. That would be perfect reliability. The second is that this should happen consistently year after year. Not only do candidates get the grades which they deserve now, but their successors must also get their just rewards.

These perceptions actually raise issues of both reliability and validity and before beginning a review of reliability alone it may be helpful to explore a little of the interaction of these two concepts, as it affects examinations.

How dependable are examination grades?

Each examination is likely to test only a sample of what might be tested; we can call this the domain of the assessment. Although a grade is gained as a result of an individual's performance on the sample chosen for a particular occasion it is widely assumed that the same grade should be obtained if other samples were used. This is partly an issue of validity: the extent to which the responses to a particular assessment can be more widely generalised.

Wiliam (1993; 1994; 1996) discusses such generalisations in terms of a four-cell model in relation to national curriculum assessment, thus:

	within domain	beyond domain
inferences	inferences with respect to the domain which is being assessed	inferences beyond the domain which is being assessed
consequences	consequences for the domain of assessment	consequences beyond the domain of assessment

The two top cells represent the whole of construct validity: within the domain in relation to the test itself and beyond the domain in relation also to the relevance and utility of the test. Clearly the inferences which may be made within the domain will depend on the extent to which the sampling of that domain is sufficiently extensive

and whether we can be confident that the measures are reliable. Although, in theory, a test might be devised which sampled the domain so completely that its validity was limited only by the reliability of the measures, this is very unlikely ever to be the case, and never to be true of public examinations, although it was an approach which was pursued with some enthusiasm in the early days of criterion-referenced testing. We are therefore interested in the adequacy of the sampling of what might be tested and in the nature of the inferences beyond what is in a particular test.

Wiliam then points out that the domain sampling for any test may have consequences beyond its construct validity. Thus, the emphasis put on examination components (whether there are closed or open response questions or how much coursework there is) or the emphasis put on certain topics or on the assessment of certain skills, go a long way to defining a subject or a syllabus and a set of values associated with these. A particular test may be used to emphasise certain values independently of considerations of reliability.

Beyond the chosen domain of assessment are the consequences which then follow for individuals, schools and others: the social consequences. Here the contrast with the upper cells of the model is clearest; whereas the inferences are open to objective analysis, independent of a value system, the consequences can only be viewed in relation to sets of values.

It is possible to see similar discussions in the literature on competence-based assessment, where judgments are made by assessors in relation to outcome statements which may be applied in a range of contexts. The outcome statements are often called performance criteria, and assessors are actually viewing evidence of performance and seeking to infer competence from it. In fact, the adequacy of the inference of competence (beyond the domain of assessment) will depend on the scope of the domain of assessment (the extent to which the range of performance contexts is covered) and the quality of the assessments which are made (their reliability). Reliability in this context has been addressed by Wolf (1993; 1995), in Beaumont (1996) and National Council for Vocational Qualifications/Scottish Vocational Education Council (1996); we discuss this further below.

It is common for discussions about the quality of the inference of competence to be conducted in terms of the sufficiency of evidence: the amount which is produced, the range of contexts in which performance is demonstrated, and the diversity of the evidence and the assessment methods used.

In all assessment environments the upper bound on the extent of the inference which may be made will be provided by the reliability of an assessment outcome. Within-domain inferences may involve different degrees of generalisation. For example, where students take the same test on two occasions the possibility for generalisation is very limited and is actually a measure of the stability of the result (it is a classical test-retest reliability). We can generalise with more confidence if two parallel forms of a test are used (involving a measure of parallel-forms reliability, where test items are matched) or where two or more tests randomly sample from the assessment domain. Our confidence in the generalisations can be discussed within a concept of

dependability, leading to methods for determining generalisability in a systematic way (Cronbach *et al*, 1972).

It has always been clear that discussions of reliability may not be conducted independently of discussions of validity. To these concepts must now be added that of dependability, and Wiliam (1993) provides a useful summary.

- Validity is the extent to which inferences within and outside the domain of assessment are warranted;
- Dependability is the extent to which inferences within the domain of assessment are warranted;
- Reliability is the extent to which inferences about the parts of the domain *actually* assessed are warranted.

The popular view that reliability also involves consistency of assessment over occasions takes us beyond a discussion of the reliability of any individual examination into an area where successive examinations may be seen as equal samplings of the same assessment domain, and therefore capable of supporting the same inferences. However, as we have seen in many comparability studies, whilst it is probably possible to ensure comparable inferences within the domain of the assessment it is much more difficult to be sure that inferences beyond the domain are comparable. Thus criticisms of and difficulties with examination comparability studies conducted over longer periods of time expose problems of the validity of the assessments rather than problems of their reliability or dependability.

Types of reliability

For a variety of reasons, perfect reliability is not going to happen. The aim must be to get as close as possible, given irreducible constraints. To realise that aim, it is necessary first to understand the sources of variation contributing to reliability, or lack of it. Some are more amenable to corrective treatment than others.

An examination is an encounter between candidates, some tasks (on question papers or in coursework) and examiners (including teachers if there is teacher assessment). This produces four sources of variation, although only the last three are sources of unreliability. They are

- between-candidate variation
- within-candidate variation
- between-examiner variation
- within-examiner variation.

As far as those managing examinations are concerned, the problems are that different examiners tend to award different marks to the same paper, that a single examiner tends to award different marks to the same paper on different occasions, and that these differences tend to be amplified by the greater freedom of response which coursework or the essay question permits. Obviously for objective tests (containing, perhaps, multiple choice questions), there is no examiner variation, which is the reason why they are more reliable than essay papers.

For their part candidates are obviously different and so between-candidate variation is to be expected (although we should note that, in a mastery testing situation this might not be the case). Between-candidate variation is not a source of unreliability; within-candidate variation is. It may arise from some characteristics of the candidate or some characteristics of the paper, or both.

Evidently candidates produce fluctuating performances, both across time and across questions or tasks. Variation across time is usually the result of how people are feeling on the day, and there is little or nothing that an examination board can do about that other than offer advice on preparation, and information on what dispensations are available in the case of illness on the day.

Variation across questions or topics or tasks is another matter. The issue here is consistency of performance within the same examination, or possibly across different occasions in the case of staged school assessment. Since any set of questions is but a sample of all the questions which might have been set, it follows that the more a candidate's individual question scores fluctuate on a particular set of questions the less reliably will the candidate's achievement be assessed. There is evidence, which we will come on to, that inconsistency of response is common among students at A level and also (especially) earlier.

Thus questions can be regarded as a source of unrelability, but because this is one source of within-candidate variation which can be addressed, we prefer to deal with it under that head. But what can an examination board do? It can make sure that there are enough questions set to provide stable estimates of achievement. In the matter of question choice it can ensure that the questions are equal in the skills which are assessed and their level of difficulty (Willmott & Hall, 1975). If it cannot achieve this then question choice must be reduced to a minimum, or taken out altogether. Exploiting choice amongst unequal questions is an obvious way in which candidates can produce uneven performances, leading to unreliable assessment. It may also lead to injustice, since the candidates are not being treated equally by the examination. Here again, structured and fixed response question formats, without choice, are at an advantage, although questions of the validity of the examination must always be considered alongside those of its reliability.

But a board can only go so far. There seem to be structural causes around the way students learn that are beyond the reach of any examining authority, the prevalence of what has been called the 'knowledge in pieces' approach to learning (di Sessa, 1988). Similarly, it has recently been widely argued that people develop expertise by the construction of mental models which are modified with widening experience

(Black & Wolf, 1990; Messick, 1982; Soden, 1993; Tomlinson & Kilner, 1992). Assessment of separate aspects of performance on different occasions or in different questions may not do justice to the whole competence nor take account of the modification of early learning by later learning. This is not to say that we will always want to assess holistically, but does say, perhaps, that some form of synoptic assessment may be better able to reflect any synthesis of the whole of what has been learned.

However, there is an obvious danger that a poorly constructed synoptic assessment might be concerned with recall of certain types of knowledge and not with the wider mental models which the individual has developed.. If we see skill development as 'learned behaviour' then it will have a variety of interlinked aspects, forming a coherent domain, no part of which may be more privileged than any other.

To summarise: there are three kinds or meanings of reliability which can be addressed operationally. These are between-examiner variation, within examiner variation and within candidate variation, including that across questions. We now look at these meanings in the context of competence-based assessment and then, later in the paper we describe appropriate methodology for isolating and estimating each source of variation.

Reliability in competence-based assessment

The emergence of what has come to be called competence-based assessment has promoted a new discussion of the classical issue of reliability. In Britain this is conducted in two contexts. One is NVQ, where occupational competence is designed to be assessed in the workplace (although much is assessed in colleges and training centres, and often uses simulations instead of 'real' workplace tasks), and the other is GNVQ, where competence in a vocational area is assessed in an educational environment. The recent review of 16-19 education and training (Dearing, 1996) emphasises that these qualifications and A and AS examinations operate within one framework, and there are various respects in which we may come to see influences of one upon the other. We therefore think that it is relevant if we discuss some aspects of reliability of competence-based assessments, and have drawn on the report produced as part of the review of NVQs, as a contribution to the report by Beaumont (1996) (University of Nottingham, 1995), and on a review of GNVQ assessment conducted by Wolf *et al* (1994). The discussion may also shed some light on the reliability of the assessment of coursework in examinations.

Assessment within GNVQs and NVQs is made using evidence generated by candidates within tasks, assignments or activities (including some written tests) in response to unit specifications. These specifications, which are broken down into elements, consist of performance criteria and range statements, variously supported by knowledge specifications (NVQs only) and evidence indicators which clarify aspects of what will be generated and assessed.

We can see direct parallels between the descriptions of examination reliability and

the reliability of assessment of these qualifications, based on outcome statements.

- Within candidate reliability
  - as a function of candidate inconsistency of performance: NVQ and GNVQ candidates may produce fluctuating performances across tasks, particularly since these will be spread over a long period of time and may be undertaken in widely different circumstances;
  - as a function of task variability: most NVQs and GNVQs are sufficiently complex for tasks to vary significantly in complexity and difficulty, perhaps leading to variations in candidate performance.

- Between examiner
  Those aspects of NVQs and GNVQs which are assessed by a local assessor (all NVQ and most GNVQ is teacher assessed) also depend on the assessor to devise the tasks and activities from which the candidate is to generate evidence of competence. He or she controls the conditions under which these are undertaken, and the details of the conduct of the assessment. Variations from assessor to assessor may arise from this and from variations in interpretation of the specifications.

- Within examiner
  As in examinations there may be inconsistency of assessor judgement over time.

Because of the complexity of GNVQs and (especially) NVQs it is easier to consider particular aspects of the reliability of their assessment rather than the reliability of the whole qualification. This is also the more appropriate approach, since the judgements about competence are made at the level of the unit and not at the level of the qualification, and may effectively be made at the sub-unit level. To date the focus of most of the work on reliability (of which there has been very little compared with that in the field of examinations) has been in two areas.

- The extent to which assessors agree on the interpretation of a given specification. With precise criteria, reliability becomes the key to the operation of an acceptable assessment system (Johnson and Blinkhorn, 1992), especially with assessment taking place in a multitude of occupational locations and often under conditions of simulation.

- The extent to which there is agreement between assessors concerning particular evidence. This also raises questions concerning the sufficiency of the evidence, making this issue as significant as that of reliability.

At the heart of much of the discussion of both validity and reliability of assessment in NVQs and GNVQs (such as by Hayer *et al*, 1994) is the range of factors which can affect the choice and circumstances of the tasks which are set and the range of factors which can affect assessors' judgements. Some of these contextual factors are environmental (such as whether what is done within an NVQ is a 'real' workplace task or a simulation), or are a function of the centre or assessor (such as the training which he or she has received), and some are linked to the candidate (such as the

relationship with the assessor and how much of the candidate's background the assessor takes into account). None of these issues is new and, in varying degrees, they affect all types of assessment (Gipps, 1994; Wood, 1991). A general discussion of the technical issues involved in the conduct of assessment in a competence-based (criterion-referenced) system is provided by Wolf (1993).

The intention with these qualifications was that the detailed assessment specifications would, by their very clarity, eliminate such extraneous context factors. This seems not to be the case, and insufficient clarity has emerged to guarantee that the assessments are reliable (Wolf, 1995). What emerges in many NVQs (but rather less, it seems, in GNVQs) is some variation in the ways in which candidates meet the evidence requirements. In particular, there are differences in the balance between what are known as 'performance evidence' (that is, evidence relating to the doing of a task) and 'knowledge evidence' (including whether the candidate can relate the knowledge to unfamiliar situations).

These aspects of the reliability of assessment are considerably influenced by the nature of the assessment centre, particularly in terms of the ways in which assessors are trained and supported in the conduct of assessment. It has been suggested that more reliable NVQ assessments are delivered by centres in larger enterprises than smaller ones, partly because the competence standards better reflect the operations of larger employers and candidates in smaller enterprises with more specialised work roles may have greater difficulty demonstrating the full range of performance evidence (Wolf, 1995). There is also the view from the care sector that NVQ standards are more appropriate to some settings than others, and that assessors and candidates have to adapt what they do in the light of local circumstances (Joint Awarding Bodies, 1994).

The extent or sufficiency of performance evidence influences directly the reliability of assessment. Raggat and Hevey (1995) showed that assessor decisions about the sufficiency of evidence incorporated a distinction between evidence satisfying formal NVQ assessment requirements, and evidence which gave confidence that the candidate was competent. The authors argued that decisions about the sufficiency of evidence cannot be made in simple mechanistic or quantitative terms; some aspects are qualitative, and involve the assessor in professional judgements which take account of a range of factors particular to each candidate.

There is considerable scope for inconsistencies to arise between different assessors' approaches to assessment. An assessment event can involve some or all of the following: observing candidates, examining work products, questioning candidates, and considering witness testimonies. The same assessment event may be used to cover more than one performance criterion in an element, or multiple performance criteria from a number of elements, or the knowledge and understanding required by more than one element. Decisions about how best to approach assessment events are made by the assessor. In the context of higher level broad skills assessment, Wolf and Silver (1995) found that assessor judgements vary in deciding why some work demonstrates competence and some does not. They found some evidence of emotive judgement and premature conclusions not supported by the information provided.

What systems exist to maintain or improve assessment reliability in NVQs and GNVQs? First of all, centres need to satisfy certain conditions before being allowed to offer either qualification. Then, assessments made by individuals are subject to internal and external verification, although these systems operate under some pressure and Peregrine *et al* (1994) found some variability in verifier judgements, and there is considerable anecdotal evidence of differences. Some of this arises because of a lack of clarity or consistency in the rules and conventions that underpin the system (Black 1994; Docking 1993; Mitchell & Cuthbert 1989), although improving this situation may be seen as a necessary though not a sufficient condition for valid and reliable assessment.

The use of well-established mechanisms, such as double assessment, audit procedures for assessment, self-help mutual support and agreement trialling can contribute to increased levels of reliability (Gipps 1995a; Wilmut, 1995). Some of these may operate internally in assessment centres or may be provided across centres in a network or by an external agent. It is widely held that assessor training is a key to valid and reliable assessment in GNVQs and NVQs. Whilst assessors have generally acknowledged the importance of this training in enabling them to function within a GNVQ or NVQ some may have had difficulty in contextualising and applying what they have learned. The direct value of systematic development of the staff involved in the assessment process is discussed by Macfarlane (1994). While acknowledging the huge variations within the system she isolates specific practices that impact on consistency and reliability in NVQ provision. Internal verification meetings with assessors were found to encourage more consistent assessment practices across assessors, particularly when focused on the interpretation of standards and the sufficiency of evidence. The use of interim assessment records (such as action plans and diaries), and formal assessor training were also noted as good practice.

Whilst reliability may be increased if more evidence of performance is assessed, and if different methods of assessment are used, the increase in cost and time needed are serious limitations, given the nature of assessment in NVQs and GNVQs. Merging assessments, so that they are based across a range of criteria or elements has been offered as a solution to this problem (Gonczi, 1991; Capey, 1995) but the effect on assessment reliability is unclear.

The alternative is standard external components as benchmarks, as examples of good practice in assessment, as moderating instruments, or simply as a means of securing more valid and reliable assessment. Such components can take a variety of forms, and are not confined to the use of written tests. There is evidence of some uncertainty regarding the effectiveness of current approaches to the assessment of knowledge and understanding, particularly at higher levels of NVQs (Peregrine *et al*, 1994), and it has been said that an external standardised test could be used to resolve some of the issues. However, although Prais (1991) outlines a case for the benefit of introducing written tests to strengthen reliability in the assessment process, others such as Steadman and Eraut (1994) are quick to point out that not all written tests (which range widely in type from essay tests to fixed response tests) are

more reliable than observation and the danger that validity will be sacrificed to reliability. More balanced discussions have been provided by Johnson *et al* (1995) for NVQs and Wilmut *et al* (1996) for core skills units in GNVQs.

Techniques by which reliability can be measured

Introduction

It is helpful to first go back to the place where the concept of reliability began, that is, with classical psychometric theory. So we first discuss, in those terms, why an assessment may be unreliable, and then what might be done in order to improve the situation. After that we focus on practical methods of measuring the reliability of an assessment, and include some discussion of methods which have gone beyond the classical approaches. This provides a background to the more detailed discussion in the next section of the report, which discusses some of the relevant research which has been done on the reliability of examinations.

What makes an assessment unreliable?

Classical reliability theory defines reliability as a ratio of true score variance to observed score variance; the difference between the two is error variance. Reliability cannot be measured directly, and is normally estimated using one of a number of methods involving successive measurements; the difficulty with these is that they do not yield comparable figures. The theory makes certain assumptions in relation to the nature of the true and error scores in relation to a pair of measurements; in brief these are that the

- sum of the error scores is zero
- error is independent of the true score
- errors on two sets of observed scores are independent of each other
- true scores on two occasions for one individual are equal
- two sets of scores have equal error variances.

It is important to recognise what does and what does not constitute error in this context. It is central to classical reliability theory that the error scores are randomly distributed and have nothing at all to do with what is being measured. There may also be bias which arises from a constant or correlated factor (such as may arise from systematically lenient marking or from differences in performance between males and females) (Hammersley, 1987). It is only possible to say whether bias will affect reliability if its context is known. For example, if only one person marks all tests, applying an equal bias to all scores the reliability is unaffected, but if he or she is one of a number of markers who apply different biases, the reliability is lowered. If the inferences within the domain are made separately for males and females the reliability for each will be higher than if a single inference is made.

The apparent simplicity of the classical model conceals some real difficulties in deciding about the forms in which bias and error occur, and what should be done about them. A more sophisticated, post-modernist, view of assessment seeks to open up a more sophisticated debate about the value systems and contexts within which a 'true score' can be said to exist, and therefore the conditions under which any test or examination can be said to be 'fair' (Gipps, 1993; 1994; 1995b). This takes us back to the earlier discussions of validity and reliability, and emphasises the danger of a single-minded pursuit of reliability.

How do we make assessments more reliable?

We have already indicated some of the things that an examination board might attempt to do in order to increase the reliability of its examinations. Although it can do little to control those aspects of within-candidate variation which result from the behaviour of the candidate it should seek to control those aspects which arise from the choice and structure of an examination paper.
It may, for example, be worth looking again at objective tests, bearing in mind the increased ingenuity and sophistication of response formats (Case & Swanson, 1993), and the fact that such tests are routinely given to graduates and to middle and senior managers. Clearly a board must seek to minimise differences within and between markers by operating suitable quality control procedures, and we discuss these later on. In terms of classical reliability theory, all of these procedures are aimed at reducing error variance as far as possible, although it will not generally be possible to eliminate it altogether.

Reduction of error variance has the effect of making the true and observed score variances more nearly equal. In practice this can be achieved by making the observed score variance as large as possible, so that it swamps the error variance, thus increasing reliability. In practical terms this will be achieved if the scores are stretched along the mark scale as widely as possible (a move which has other benefits in an examinations environment). This is made easier within a single paper if the item or question inter-correlations are as large as possible - something that may be more easily achieved if the assessment domain is structured around a smaller range of skills, topics or concepts.

Whereas the first strategy (of reducing errors in the conduct of the assessment process) will benefit all forms of assessment, the second (extending the observed score mark scale) produces difficulties when we are dealing with assessment based on criteria. The wish to maximise observed score variance is quite at odds with the expectation that candidates will demonstrate mastery in relation to one or more criteria. In fact, the score distribution in a mastery test is of little direct interest in terms of differentiating between candidates, except at the point where a decision has to be made about mastery or non-mastery, competence or non-competence.

Thus, apart from any difficulties which may arise with the basic assumptions of classical reliability theory (Seddon, 1988) there is the additional difficulty that the

worth of a mastery or criterion-referenced test may be seriously misrepresented if classical reliability estimates are applied to it.

Lengthening an examination

Classical reliability theory enables one to show that a longer test will be more reliable than a shorter one, provided that the additional items or questions are drawn from the same universe on the same basis as those originally in the test (Ebel, 1972). It may also be true that a test which comprises a large number of short questions, representing a domain of assessment, will be both more valid and more reliable than a test which comprises a small number of long questions, particularly when there is a choice of questions.

It is less clear whether the addition of a component to an examination will increase the reliability of that examination. Such a move might be made on the grounds of validity, and the effect on the examination as a whole would undoubtedly depend on the weighting given to the new component and the reliability of the component itself. Cresswell (1984) has shown the percentage weightings which may be allocated to a new component of varying reliability and with different correlations with the whole examination. With low weighting for the new component the effect on overall reliability may be negligible; with high weighting the reliability of the new component would be a crucial factor.

The scope for examination lengthening is, of course, severely limited by a variety of practical considerations, and the procedure is unlikely to be an acceptable (or effective) way of solving the problems of reliability of an indifferent test. The Task Group on Assessment and Testing (TGAT) report suggests the need for a balance to "... minimise the amount of information gathered while maximising the confidence in its interpretation" (Department of Education and Science/Welsh Office, 1987); this must be a goal for all examination systems.

Some approaches to measuring reliability

Classical approaches to getting a measure of the reliability of an examination are to repeat it or to operate it in two parallel forms, but these methods are not to be contemplated in practice (this one of those 'irreducible constraints'). The determination of the reliability of the grades cannot be approached directly, and we must break down an examination into its various parts and processes and work on the reliability of each. We will, in practice, work on the sources of variation which are susceptible to investigation, on the basis that the more that is done to tighten up in each of these areas the greater will be the confidence in the grades awarded.

Within-candidate variation is addressed by calculating the internal consistency of tests and papers. This is common practice for dichotomously scored multiple choice tests, but the method has been generalised to polychotomously scored essay papers (Backhouse, 1972). The same method even incorporates a variation which takes into

account question choice, and a very substantial discussion is provided by Willmott & Hall (1975). In calculating internal consistency it is necessary to be aware of statistical artifacts which may come into play, especially the effect of many individuals omitting or not reaching questions. This gives a false impression of consistency, and spuriously drives up the reliability estimate.

There is a range of internal consistency measures which are generally not exactly comparable, but which may be used for routine monitoring of any one question paper. They are discussed in detail by Willmott & Nuttall (1975). From them estimates may be made of the reliability of a whole examination. Such estimates are probably conservative.

Examinations are then graded, and the partitioning of the final mark scale into grades reduces the reliability of the examination as a whole, since there is a loss of information. This results in some candidates getting a grade which is one or more higher or one or more lower than they should get. It is not possible to determine exactly how many candidates are thus affected, since we do not know what grade a candidate ought to obtain. However, work with simulated distributions and different reliabilities for the total mark scale enable us to anticipate the degree of mis-classification which would probably result. As an example, Cresswell (1986) showed that the partitioning of a set of marks having a reliability of 0.90 into 7 grades would result in an overall reliability of 0.77 and that a quarter of candidates would get an 'incorrect' grade. The reliability of the levels rises as the number increases, that is, gets closer to the number of mark points in the mark scale. Although there is a corresponding decrease in the proportion of candidates getting the correct level, each misclassification is less serious if there is a large number of levels, rather than a small number. However, some of the misclassifications may be by more than one level, when there are many of them, and exact outcomes depend to some extent on the shape of the mark distribution and, particularly, the assumed proportions of candidates to be placed in each of the levels.

Between and within examiner variation can be addressed by having examiners re-mark scripts, both their own (so as to get an estimate of within-examiner variation), and others (to get an estimate of between-examiner variation). Random selections of scripts are required, and all scripts must have previous marks and examiners' annotations removed (Murphy, 1979). There is no need to resort to copies, which removes any worry that examiners may treat copies differently from originals (Braun, 1988), although there have been many small studies using copied scripts where this problem has not been identified. Additionally, in a small-scale study, Wilmut (1984) compared the marking of copied scripts with all marks and annotations removed, with marks only removed, and with marks and annotations present, and found no differences in marker decisions, and Newton (1996) found no difference between two samples of scripts used in a re-marking exercise, where one sample had only the marks removed and the other had both marks and annotations removed.

The question for examination boards is what to do with the findings from such exercises. Some of the issues which arise are concerned with the conduct of the

marking process (how scripts are distributed, what information is on the scripts, how examiners' work is monitored and what adjustments might be made to compensate for errant marking), and some are concerned with the construction and use of marking schemes. We look briefly at both of these areas in the next section.

As regards teacher assessment, it need not be treated any differently: there can be within and across school re-marking exercises, from which reliabilities can be estimated. There is a brief report on such work in the next section. However, expectations appear to have changed somewhat since the first GCSE General Criteria (Department of Education and Science, 1985) emerged. There it was said that teachers' mark rank orders would remain unaltered and only scalings would be applied in order to achieve parity of standards. Such adjustments do not, of course, alter the reliability of any given set of marks, although they would (if appropriately chosen) improve the reliability of marks across many teachers, by eliminating at least some aspects of bias. The GCSE Mandatory Code of Practice (School Curriculum and Assessment Authority, 1995), as part of a detailed set of procedures for the conduct and moderation of teacher assessment, allows for rank orders to be changed where '.. a centre has been demonstrably inconsistent'. In such circumstances it is impossible to say whether the reliability of the resulting marks is raised or lowered.

A better methodology

Mounting a lot of exercises to pick up the different kinds of reliability is fine as far as it goes, but it would be altogether cleaner and more satisfying if a methodology was available which took into account all identifiable and estimable sources of variation simultaneously. Such an integrated methodology is available: it is called generalisability theory (although 'theory' is a bit of a misnomer, and 'analysis' would be preferable). It replaces what has been called the 'one source of error at a time' approach of classical psychometric theory (Swanson *et al*, 1987) with an integrated framework and analytical routines which enable the user to evaluate the influence of multiple sources of variation within the desired universe of inference (Cronbach *et al*, 1972). The greatest gain is a superior conceptualisation of reliability which ties it directly to the intended use of scores or grades.

Generalisability analysis provides the statistical apparatus for answering the fundamental question: 'given a candidate's performance on a particular task at a particular point in time assessed by a particular assessor, how dependable is the inference about how that individual would have performed across all occasions, tasks, observers and settings?' To estimate dependability, the individual's performance needs to be observed on a sample of tasks, on different occasions, in different settings, and with different observers. In the examinations context, the selection of tasks would be the questions or papers (you could work at either level). The different settings might mean examination hall and classroom, and, as for occasions, these could be the different times at which school assessments are made (there is one study of this in the literature by Wood, 1976).

The methodology has the additional advantage that it can address within-candidate variation over time. With any group of candidates it is not possible to keep the tasks the same on successive testing occasions. We either have to contend with the learning which has taken place as a result of the first testing occasion (although we may be able to say that this is small if the testing occasions are a long way apart, unless other learning experiences intervene to produce unpredictable effects), or the two testing occasions must use different materials. In this case differences between the tests and differences between the candidates become confounded, even though the two tests may be equal samples from the same universe of possible examinations. Here again generalisability analysis has much more to offer than classical reliability theory, since it enables us to identify all of the sources of variation within a single model. One such study (though not in the field of examinations as we have them in Britain) was reported by Cardinet *et al* (1976). There is an additional benefit that the methodology of generalisability anlaysis provides us with a link between work on reliability and work on comparability of standards.

The other special beauty of generalisability analysis is that once the important sources of variation have been isolated and quantified it becomes possible to start forecasting what levels of reliability might be obtainable by tweaking the system. For example, what happens when you increase the number of times a question is marked, or increase the number of questions on a paper, or decrease the number of papers?

Generalisability analysis is recommended as the preferred methodology for investigating examination reliability, and we have included some outline proposals in the last section of this paper. There has been a major published application of this methodology to British examinations data by Johnson and Cohen (1983, 1984), who applied it to one Board's comparability studies, and were able to say something useful about the lack of reliability of O level French as exhibited in the inability of examiners to agree on appropriate marks. In addition there is a significant note in the Task Group on Assessment and Testing (TGAT) report (as Appendix G: Johnson, 1987), referring to the use of the method in the various interpretations of national curriculum assessment outcomes. This is based on its use within the Assesment of Performance Unit science studies, and the suggestions in the note exactly parallel the flexibility referred to in the previous paragraph.

Research into examination reliability

Introduction

The descriptions given in this section are designed to support the main discussion of earlier sections; by and large, these are not repeated here. Some pieces of work have already been referred to (and are listed in the References) and some further evidence is available from the marking reliability bibliography.

An extended account of relevant research up to 1991 can be found in Wood's survey, whose scope extends beyond British examinations (Wood, 1991).

Studies of between-examiner variation

At the outset it is important to note that, although there have been quite a few empirical investigations of inter-marker agreement, and that these are relatively easy to conduct and yield immediately useful results, examination reliability 'is a much more complex issue than this' (Johnson, 1988). This type of work goes back through the whole of the twentieth century; we will look only at a selection of more recent studies.

Re-mark exercises were carried out by Murphy at the end of the 1970s (Murphy, 1978, 1982). He noted that the least reliably marked examinations tended to be those that place the most dependence on essay-type questions, and the most reliably marked examinations tended to be those that are made up of highly structured, analytically marked questions.

Concerning marker behaviour, an American study, now 30 years old (Godshalk *et al*, 1966) arrived at the conclusion, which has never been challenged, that the reliability of essay scores is primarily a function of the number of different topics to be tackled and the number of readings (markings) that the essays got. The increases which can be achieved by adding topics or markers are dramatically greater, the authors thought, than those which can be achieved by lengthening the time per topic or developing special procedures for marking.

Moving from single marking to double marking would definitely be an improvement but maybe a move to triple marking, for instance, would produce diminishing returns. That is what Wood and Quinn (1976) found. They also demonstrated that there is nothing to be gained from attempting to pair examiners according to known or suspected marking characteristics rather than in some random or semi-random way.

It is not thought that any school examination boards go in routinely for double marking arrangements on any large scale. The logistics would certainly be

formidable. The Council of Legal Education, which operates an annual selection process for choosing students to go on the Bar Vocational Course has shown for the past two years that it is possible to operate double marking on three written papers, and to have the marking done blind (Wood *et al*, 1996). The entries are small (only about 1500 or so) but it does show what can be done when the will is there to achieve equitable treatment all round.

There is evidence that some kinds of writing can be assessed more reliably than others. This fits with the clear research finding that markers are quite unable to distinguish (at least when marking) between different features of writing (Wood, 1991). This also emerged from the study of Key Stage 3 assessment of English in 1995 (University of Exeter, 1995), where markers failed to make distinctions between the mechanics of writing and the candidates' capacity to demonstrate understanding and write expressively; they generally failed to reward the latter. It is, of course, the case that the most reliable aspects of the marking in both studies will have been concerned with the mechanics, and this suggests that the prospects for analytic marking may not be good.

In fact, mark schemes have become progressively more structured (though not necessarily more analytical) over the last few years, and the overall quality of marking is certainly much higher than it once was. Wilmut (1982) suggested that it was possible that we had then reached the end of this progression and must look to a more detailed analysis of marking variability in order further to reduce inconsistencies. Some of the factors affecting consistency described by Wade (1978), Husbands (1976), Alston (1983) and Branthwaite, Trueman & Berrisford (1981), such as sex of the candidate, handwriting, ideological stance of the examiner or aspects of examiner personality, may be of less importance in the examination context than some factors mentioned by Dunstan (1966), such as speed of reading, tiredness and the academic competence of markers. Limited evidence presented by Wilmut suggested that thoroughness of marking is a major issue, and a study by Breland and Jones (1988) suggested that greater consistency of marking can be achieved when markers work in teams (a 'conference' setting) than when they work singly, even when monitored.

Examination boards are therefore as likely to be concerned with random variations in marker behaviour as they are to be concerned with systematic bias and linear adjustments to cope with severity or leniency are often irrelevant (Spencer, 1981). However there may still be a case for removing as many potential sources of bias as possible. If the boards will not entertain double marking then they might at least send out random samples of scripts to examiners. This would stop examiners getting disproportionate numbers of scripts from the same kind of centres, or perhaps only ever seeing scripts from one kind of centre. Calibre of scripts seen is bound to influence marking. There is another benefit. Were scripts to be randomised then, because of the randomisation, the board could constrain examiners to supply fixed percentages of candidates at each mark. Within a monitoring programme, sets of scripts at each mark ought then to be found to be comparable (Wood, 1991), and if they are, reliability would be more assured.

The removal of names and school identifiers from scripts would also ensure the avoidance of bias; the case for this is argued by Fitz-Gibbon (1996).

Within examiner variation

Little is known about within-examiner variation. There is the rather depressing finding that the agreement between the same examiner's mark on two different occasions was scarcely better than between two different markers (Wood, 1991). Also, in a certain comparability exercise there was as much variation on grading standards amongst examiners from the same board as there was among examiners from different boards (Johnson, 1988). There are also some pointers in the literature suggesting increased variability in marking standards as marking progresses (Wood, 1991); not surprising really, since the more scripts you see the more the true calibre of the entry is revealed to you, and the more you adjust as you go along.

Within candidate variation

Concerning within-candidate variation it was noted that inconsistency of response has been observed. Johnson saw it in A level mathematics where it was quite common for candidates to show discrepant performances within and between the Pure and Applied papers (Johnson, 1988). In seeking to explain what was going on Wood (1991) commented on the tendency for syllabuses to become one huge cafeteria in which students (and their teachers) are given carte blanche to pick and mix. With fragmentation dominant, instead of developmental progression, the result is inevitably an achievement texture which is bitty and incoherent. Murphy (1988) wrote of the inability of students to generalise across situations or else their propensity to apply the same problem-solving strategy irrespective of the problem.

School-based assessments

Little has been published on the reliability of school-based assessments since the 1967 Joint Matriculation Board study (Hewitt, 1967) although there have been plenty of commentators willing to claim that teacher assessment must be hopelessly biased and cannot be trusted. As it happens, the JMB study reported an average correlation of 0.83 between the school's assessment and the assessment of an independent moderator (for 20 candidates in each of 10 schools). Not one of the correlations fell below 0.60. This compares very favourably with what might be expected from any two examiners marking an essay paper.

The growth of CSE examinations in the late 1960s generated a considerable amount of interest in the use of consensus moderation and agreement trialling. One study from that period compared teachers' judgements in English, Mathematics, Chemistry and History (Cohen, 1974), and worked with both scripts and coursework. With English scripts, differences between assessors on a single occasion were not as large as some inconsistencies within assessors between occasions and, whereas the inter-

assessor reliability with History scripts was 0.92 (comparing well with the Chemistry reliability of 0.95), with coursework it was only 0.61.

Willmott & Nuttall (1975) did not attempt to determine reliabilities for teacher assessed components within the examinations which they studied, and 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, what they called the 'subjective nature' of the assessment and the many teachers involved would offset this gain. Much more recently Taylor (1992) has reported very creditable correlations in the region of 0.87 to 0.97 between pairs of moderators marking coursework folders in English and Mathematics. If the teachers for whom they are responsible operate similarly then the results of the earlier JMB study appear to be borne out.

What we do not have evidence for is the nature and extent of bias in teacher assessment, paralleling that which we have already discussed for NVQs and GNVQs (where it appears to be a function of centre and assessor variables, and assessor-candidate interactions). 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). We do not say that this happens but if the model for a fair and equitable assessment process is something like double-marking-blind then evidently school-based assessment may fall some way short.

Generalisability analysis

The two applications of generalisability analysis to British examinations have already been noted, together with some other relevant work. Otherwise there is a worked example of a related application using the relevant software which shows how it all works (Wood *et al*, 1989).

Applications of the methodology in a medical examining context have been reported in recent times. One shows how to use multivariate generalisability analysis to estimate the reliability of an examination treated as a composite and the contributions of each component (Hays, 1995). The other reports that the major source of unreliability is not between examiner variation (because examiners can always be trained) but rather within candidate variation, which is interesting, given what we have been saying earlier in this paper about the potential for uneven performance across tasks (van der Vleuten & Swanson, 1990). The paper stresses the absolute necessity of providing sufficiently large numbers of assessment tasks in order to obtain stable, reproducible assessment of examinee skills.

Recommendations for further work

As far as we know the examination boards are continuing routine monitoring of aspects of reliability, in relation, for example, to consistency of marking of GCSE and A level. We assume that SCAA is able to monitor the results of this work, if it does not already do so, and links it with findings from the monitoring of national curriculum assessment and competency-based assessment. We have therefore concentrated on some of the other more general issues which we consider would be worth pursuing. They are not all matters of research.

- **Enquiries into marking behaviour**
  The evidence reported here indicates that we still have a great deal to learn about the interactions between examiner and mark scheme and examiner and script, and the factors which may lead to what is seen as error variance, but which may turn out to be forms of bias. Systematic and fundamental research in this area, in the context of modern examinations (including, perhaps, national curriculum assessment) is desirable. Possible studies include

  - a detailed study of the processes of marker standardisation; whether these are effective; the extent to which markers share a common understanding of the interpretation of a mark scheme

  - a systematic classification of marker 'errors': points at which markers differ from one another or interpret the mark scheme in an incorrect of idiosyncratic fashion

  - an examination of the effectiveness of marker monitoring procedures and the credibility of post-marking adjustment procedures and decisions.

  Such studies may be of a survey form, but may be experimental; important underpinning features would be interviews with markers, and the detailed tracking of marker behaviour when dealing with operational scripts. In some subjects it may be important to distinguish between marking at GCSE and at A level, and the marking and moderation of coursework may be treated in a similar fashion to script marking.

- **Removing marker bias**
  Linked to the above is the need to examine whether it is feasible to remove some of the factors which may cause bias in marking, such as marker knowledge of the candidate or centre. Bias may arise in relation to gender, ethnicity, social class or religion, based on assumptions arising from the candidate's name, and from the name and location of the school or college. There may here be matters of discussion with the boards, but the issues could also be the subject of research studies. Areas for experimentation might include

  - the removal of centre and candidate identification, other than numbers, from

scripts and coursework

- random script allocation amongst markers.

- **The reliability of school-based assessments**
  A broadly-based investigation of the reliability of school-based assessments is required, perhaps building on the earlier CSE and JMB studies, and on more recent work. Because of the broadening experience with GNVQs and NVQs this may be extended beyond the current limited scope of coursework assessment for public examinations, and might include studies of the appropriateness of different models of moderation and verification. For example,

  - studies of the effectiveness of verification procedures which are based on centre accreditation and on checks that assessment processes (including internal moderation) have been carried out satisfactorily; are these more or less effective than moderation procedures, based on sample re-marking?

  - the mechanisms which centres use for internal standardisation and moderation; how effective these are, and their effect on the reliability of coursework assessment

  - a review of the acceptability and effectiveness of assessment training and of assessment support materials in ensuring valid and reliable coursework assessment.

- **Analysis of entire examinations**
  Further work using generalisability analysis would be timely, and should be linked to work on comparability. Such work has the capability of acting as a vehicle for a most thorough and systematic exploration of reliability and related issues, with a scope which can range from the study of a single examination to a study of all examinations (if one had the means). Because of its capacity for treating variables in a flexible way it can also enable some of the other work in this list. For example, it could simultaneously allow the exploration of different types of mark schemes, variations of marking behaviour over time, the effects of random script allocation, and so on.

  Whilst grand designs are certainly possible in this area, a range of smaller studies may also be appropriate, focusing on different aspects of examinations. Their disadvantage is the loss of flexibility and the much more limited scope for exploring alternatives.

- **Test formats**
  It may be appropriate and timely to conduct an investigation of the greater use of objective tests in GCSE (and, perhaps, A level) examinations, bearing in mind the increased ingenuity and sophistication of response formats, the fact that such tests are routinely given to graduates and to middle and senior managers, and the apparent capacity to assess a wider range of skills and achievements.

  It may also be appropriate to look at the very flexible test formats which appear to be emerging in national curriculum tests at Key Stages 2 and 3, and to see the extent to which these should be reflected in some parts of GCSE and A level practice.

- **Broader studies of reliability**
  Studies of examination reliability which are conducted only in the context of GCSE and A level examinations may be less valuable than those which look more widely across the spectrum of assessment. For example, it may now be increasingly appropriate to study the reliability of assessment in more detail across GNVQs and, perhaps, NVQs, as a contribution to the debate about the appropriateness of various patterns of internal and external assessment in all of these qualifications.

  Moreover, with the increasing emphasis on the integration of educational provision 14-19 and on more flexible progression into higher education (perhaps with the increased availability of credit which can be used for HE programmes), it may be timely to study in more detail the relationships between various technical requirements (including reliability) for assessment at all of these levels.

References

ALSTON, J. (1983) A Legibility Index: can handwriting be measured? *Educational Review*. 35.3 237-242

BACKHOUSE, J. (1972) 'Reliability of GCE examinations: A theoretical and empirical approach' in NUTTALL, D.L. & WILLMOTT, A.S. *British Examinations: Techniques of Analysis.* Slough: NFER Publishing Co.

BEAUMONT, G. (1996) *Review of 100 NVQs and SVQs: A report submitted to the Department for Education and Employment.*

BLACK, H. (1994), Sufficiency of evidence, *Competency and Assessment* (20), 3-9

BLACK, P. AND WOLF, A. (1990) *Knowledge and competence: current issues in training and education*. Sheffield: COIC

BRANTHWAITE, A., TRUEMAN, M. & BERRISFORD, T. (1981) Unreliability of Marking: further evidence and a possible explanation. *Educational Review* 33(1) 41-46

BRAUN, H.I. (1988) Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics.* 13.1-18

BRELAND, H.M. & JONES, R.J. (1988) *Remote Scoring of Essays.* College Board Report 88-3. New York: College Entrance Examination Board

CAPEY, J. (1995) *GNVQ Assessment Review*. London: National Council for Vocational Qualifications

CARDINET, J., TOURNER, Y. & ALLAL, L. (1976) The Symmetry of Generalisability Theory: Applications to Educational Measurement. *Journal of Educational Measurement*. 13(2) 183-204

CASE, S.M. & SWANSON, D.B. (1993) Extended matching items: a practical alternative to free response questions. *Teaching and Learning in Medicine.* 5, 107-115

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

CRESSWELL, M.J. (1984) *The effect on reliability of adding an additiona component to an examination*. AEB Research Paper RAC 333

CRESSWELL, M.J. (1986) How many examination grades should there be? *British Educational Research Journal* 12.1. 37-54

CRONBACH, L.J., GLESER, G.C., NANDA, H. & RAJARATNAM, N. (1972) *The Dependability of Behavioral Measurements: Theory of generalizability for scores and profiles.* New York: Wiley

DEPARTMENT OF EDUCATION AND SCIENCE (1985) *GCSE General Criteria.* London: HMSO

DEPARTMENT OF EDUCATION AND SCIENCE/WELSH OFFICE (1987) *Report of the Task Group on Assessment and Testing.* London: HMSO

DI SESSA, A. (1988) 'Knowledge in pieces' in FORMAN, G. & PUFALL, P.B. (eds) *Constructivism in the Computer Age.* Hillsdale, NJ: Lawrence Erlbaum Associates. pp49-70.

DOCKING, R.A. (1993), Assessment in the workplace: facts and fallacies, *Competence and Assessment* (15)

DUNSTAN, M. (1966) 'Sources of Variations in Examination Marks' in HEYWOOD, J. & ILIFFE, A.H. (eds.) *Some Problems of Testing Academic Performance.* Dept. of Higher Education Bulletin 1. Lancaster: University of Lancaster

EBEL, R.L. (1972) Why is a longer test usually a more reliable test?, *Educational and Psychological Measurement*, 32, 249-253

FITZ-GIBBON, C.T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness*. London: Cassell

GIPPS, C.V. (1993) The Profession of Educational Research. *British Educational Research Journal* 19.1 3-16

GIPPS, C.V. (1994) *Beyond testing: towards a theory of educational assessment*, Falmer Press, London & Washington D.C.

GIPPS, C.V. (1995a) 'Reliability, validity and manageability in large scale performance assessment' in Torrance, H. (ed) *Evaluating Authentic Assessment* Open University Press, Buckingham

GIPPS, C.V. (1995b) What Do We Mean by Equity in Relation to Assessment? *Assessment in Education*. 2.3 271-282.

GODSHALK, F., SWINEFORD, F. & COFFMAN, W.E. (1966) *The measurement of writing ability*. College Board Research Monographs. No. 6

GONCZI, A. (1991) Competency based assessment in the professions in Australia, *Assessment in Education* i 1, 27-44

HAMMERSLEY, M. (1987) Some notes on the terms 'validity' and 'reliability'. *British Educational Research Journal.* 13(1). 73-82

HAYER, P., GONCZI, A. & ATHANASOU, J. (1994), General issues about assessment of competence, *Assessment and Evaluation in Higher Education* xix 1, 3-16

HAYS, R.B. *et al* (1995) Longitudinal reliability of the Royal Australian College of General Practitioners Certificate Examination. *Medical Education*. 29, 317-321

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

HUSBANDS, C.T. (1976) Ideological bias in the marking of examinations: a method of testing for its presence and its implications. *Research in Education*. 15. 17-38

JOHNSON, S. (1987) 'Technical note on reliability, validity and error' in DEPARTMENT OF EDUCATION AND SCIENCE/WELSH OFFICE *Report of the Task Group on Assessment and Testing*. London: HMSO

JOHNSON, S. (1988) Comparability in degree awards: implications of two decades of secondary level research. *Studies in Higher Education*. 13, 177-197

JOHNSON, S. & COHEN, L. (1983) *Investigating Grade Comparability through Cross-Moderation*. London: Schools Council

JOHNSON, S. & COHEN, L. (1984) Cross-Moderation: a useful comparative technique. *British Educational Research Journal*. 10, 89-97

JOHNSON, C. & BLINKHORN, S. (1992), *Validating NVQ assessment* Research and Development Series, Report No. 7, Department of Employment, Sheffield

JOHNSON, C., WOLF, A. & BARTRAM, D. (1995) *External Assessment for NVQs.* Report to NCVQ for the Review of 100 NVQs/SVQs

JOINT AWARDING BODIES (1994), *Assessment in NVQs in the Care Sector* London: Joint Awarding Bodies

MACFARLANE, K. (1994), Towards best assessment practice, *Competence and*

Assessment (22)
MESSICK, S. (1982) *Abilities and knowledge in educational achievement testing: the assessment of dynamic cognitive structures.* Princeton: Educational Testing Service
MITCHELL, L & CUTHBERT, T. (1989), *Insufficient Evidence: The Final Report of the Competency Testing Project*, SCOTVEC, Glasgow
MURPHY, P. (1988) Insights into pupils' responses to practical investigations for the APU. *Physics Education*. 23, 330-336
MURPHY, R.J.L. (1978) Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*. 48, 196-200
MURPHY, R.J.L. (1979) Removing the marks from examination scripts before re-marking them: does it make any difference? *British Journal of Educational Psychology*. 49. 73-78
MURPHY, R.J.L. (1982) A further report of investigations into the reliability of GCE examinations. *British Journal of Educational Psychology*. 52, 58-63
NATIONAL COUNCIL FOR VOCATIONAL QUALIFICATIONS/SCOTTISH VOCATIONAL EDUCATION COUNCIL (1996) *Review of 100 NVQs/SVQs: A report of the findings*. London: NCVQ
NEWTON, P.E. (1996) The reliability of marking of GCSE scripts: Mathematics and English. *British Educational Research Journal* (to appear)
NUTTALL, D.L. (1987) The validity of assessments. *European Journal of Psychology of Education*. 2(2). 109-118
NUTTALL, D.L. & WILLMOTT, A.S. (1972) *British Examinations: Techniques of Analysis*. Slough: NFER Publishing Co.
PEREGRINE, P., PEDRESCHI, T., CONNOR, J., THACKRAY, D. & WOLSTENCROFT, T. (1994), *Effective practice in assessment against the management standards* Research and Development Series, Report No. 24, Department of Employment, Sheffield
PRAIS, S. (1991), Vocational qualifications in Britain and Europe: theory and practice, *National Institute Economics Review*, 86-92
RAGGAT, P. & HEVEY, D. (1995), *Sufficiency of evidence* Research and Development Series, Report No. 32, Department of Employment, Sheffield
SCAA (1995) *GCSE Mandatory Code of Practice*. London: School Curriculum and Assessment Authority
SEDDON, G.M. (1988) The validity of reliability measures. *British Educational Research Journal*. 14(1) 89-98
SPENCER, E. (1981) Inter-marker unreliability in GCE Ordinary grade English Composition. Is Improvement Possible? *Scottish Educational Review* 13(1) 44-55
SODEN, R. (1993) *Teaching thinking skills in vocational education*. Sheffield: Employment Department
STEADMAN, S. & ERAUT, M. (1994), *Can you validly put your trust in reliability? Sig Prais and NVQs*. MS from University of Sussex School of Education
SWANSON, D.B., NOREINI, J.J. & GROSSO, I.J. (1987) Assessment of clinical competence: written and computer based simulations. *Assessment and Evaluation in Higher Education*, 12, 220-246
TAYLOR, M. (1992) *The reliability of judgements made by coursework assessors*. AEB Research Report RAC 577.
TOMLINSON, R. AND KILNER, S. (1992) *Flexible Learning, Flexible Teaching: the*

flexible learning framework and current educational theory. Sheffield: Employment Department

UNIVERSITY OF EXETER (1995) *Evaluation of the quality of external marking of the 1995 Key Stage 3 tests in English*. Report to SCAA

UNIVERSITY OF NOTTINGHAM (1995) *The Reliability of Assessment of NVQs*. Report to NCVQ.

VAN DER VLEUTEN, C.F.M. & SWANSON, D.B. (1990) Assessment of clinical skills with standardised patients: state of the art. *Teaching and Learning in Medicine*. 2, 58-76

WADE, B. (1978) Responses to written work. *Educational Review*. 30. 149-158

WILIAM, D. (1993) Validity, dependability and reliability in National Curriculum assessment. *The Curriculum Journal* 4(3) 335-350

WILIAM, D. (1994) 'Reconceptualising validity, dependability and reliability for National Curriculum Assessment' in HUTCHISON, D. & SCHAGEN, I. *How Reliable is National Curriculum Assessment?* Slough: National Foundation for Educational Research.

WILIAM, D. (1996) National Curriculum Assessments and Programmes of Study: validity and impact. *British Educational Research Journal* 22(1). 129-142

WILLMOTT, A.S. & HALL, C.G.W. (1975) *O Level examined: the Effect of Question Choice*. London: Macmillan Education

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

WILMUT, J. (1982) *Marking examination scripts: studies of differences amongst examiners*. Paper presented at the conference of the British Educational Research Association, St. Andrews.

WILMUT, J. (1984) *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC 315.

WILMUT, J. (1995), *Agreement Trialling for Professional Development in Assessment* Paper presented to the 21st annual conference of the International Association for Educational Assessment, Montreal, Canada, June 1995

WILMUT, J., MACINTOSH, H. & WOOD, R. (1996) *The feasibility of some external assessment of the core skill units in GNVQ*. Report to NCVQ.

WOLF, A. (1993), *Assessment Issues and Problems in a Criterion Based System* Further Education Unit, London

WOLF, A. (1995), *Competence-Based Assessment* Open University Press, Buckingham

WOLF, A. & SILVER, R. (1995), *Measuring `broad' skills: design issues for assessment* Research and Development Series Report No. 31, Department of Employment, Sheffield.

WOLF, A., BURGESS, R., STOTT, H. & VEASEY, J. (1994) *GNVQ Assessment Review Project*. London: University of London Institute of Education

WOOD, R. (1976) Halo and other effects in teacher assessment. *Durham Research Review* 7, 1120-1126

WOOD, R. (1991) *Assessment and Testing: A Survey of Research*. Cambridge: Cambridge University Press

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

Studies. 19(5) 409-424

WOOD, R. & QUINN, B. (1976) Double impression marking of English Language essay and summary questions. *Educational Review*. 28. 229-246

WOOD, R., HAMER, G., JOHNSON, C.E. & PAYNE, T. (1996) 'Selection for the Bar: results of the last three selection processes'. In ANDERSON, N. & HERRIOT, P. (eds) *Handbook of Selection and Appraisal* (2nd ed). Chichester: John Wiley (to appear)

WOOD, R., JOHNSON, C.E., BLINKHORN, S,F., ANDERSON, S.A. & HALL, J.P. (1989) *Boning, Blanching and Backtacking: Assessing Performance in the Workplace.* Sheffield: Training Agency