Applications of process control for the maintenance of standards and the quality assurance of results in a connoisseurship model of assessment.

Paper to be presented at the 33rd IAEA Conference,
Baku, September 2007

Author and presenter:
Dr. Jonathan H Robbins
Correspondence to:
The Talent Centre Nottington Weymouth Dorset DT3 4BH England +44 (0)1305 814820

www.talent-centre.com

Abstract

Classical methods of estimating reliability based on correlations are not appropriate for a connoisseurship model of assessment. Difficulties arise because of the complexity of the judgements that have to be made, the extent to which inferences are drawn and the assumptions on which these are based. In a connoisseurship model of assessment the meaning of reliability depends on a complex and often unpredictable mix of context, performance variables, the quality of assessor decisions and external factors. These issues are discussed and a rationale for redefining reliability as the stability of judgements is provided; the basis for this is described in relation to work in the fields of assessments of performance by airline cockpit crews and forms of statistical process control in precision engineering. Applications of the concept of stability of judgements in assessments of performance are considered and illustrated with examples using software developed for use by awarding bodies. Issues relating to quality assurance, the maintenance of standards and forms of reporting are considered and some conclusions drawn about applications, outcomes, and future developments.

A connoisseurship model of assessment

In the United Kingdom the use of a connoisseurship model of assessment is widespread and has a long tradition, especially in the arts and for teacher assessment of coursework, portfolios and investigative or project work. The extent of this practice and the assumptions associated with it, leads to a need to clarify what is meant by the term connoisseurship when applied to assessment. This is important for two reasons, firstly because it is easy to assume that what is implied is a sort of appreciation or valuing that lacks the rigour thought to exist in other sorts of assessment but which nonetheless has value, particularly to the student. Secondly, because the credibility of the judgements made, depend on the status and standing of the connoisseur, both in relation to a community of practice and on the extent to which this particular community of practice is acknowledged and esteemed by those who form the social context in which it operates. This last point is particularly relevant in a society in which the authority of office, position or role is not simply accepted or deferred to, but must continually be justified by inspection or proof of value.

The Oxford English Dictionary defines a connoisseur as “one aesthetically versed in any subject, esp., one who understands the details, technique, or principles of a fine art; one competent to act as a critical judge of an art, or in matters of taste (e.g. of wines etc.)”. Three characteristics of a connoisseur may be inferred from this, (i) the person is qualified to do so, (ii) the exercise of critical faculties is based on knowledge and (iii) an ability to make comparisons in relation to perceived qualities. The term ‘educational connoisseurship’ is used by Eisner (1998a) to mean an art of appreciation arising from expertise in the domain of education and educational criticism as the art of and the vehicle for disclosure of judgements to a wider audience (For an exploration of his thinking about this see, for example, Eisner (1985) and (1998b). Stating that: “Educational connoisseurship is the art of appreciation. Educational criticism is the art of disclosure.” Eisner describes connoisseurs as people who enjoy and understand and critics as “people who transform the contents of connoisseurship into a public language that makes it possible for others less sophisticated in that particular domain, to notice the qualities that critic writes about.” Eisner puts forward the view that that anyone involved in education has the right and responsibility to be a critic, but that certain people must be trained in order for an authentic connoisseurship to be exercised. Despite being frequently referred to the assessment literature, the practice of educational connoisseurship is not something that appears to have been widely accepted in the United States. Part of the reason for this non-acceptance is a cultural one that arises from the grading of students and the uses of standardised testing and surfaces in the debates between proponents of authentic assessment and their opponents. However part of the reason (and one that has similarities with the general acceptance of teacher assessment in the United Kingdom), must also be the extent to which the authority and purpose of the person acting as connoisseur and critic, is accepted by those involved. If the wider community does not accept the purpose of a critique, however well informed by connoisseurship, then the pronouncements of the critic have no authority.

The writings of Michael Polanyi (1958) provide valuable understandings in relation to connoisseurship, particularly in relation to notions of knowledge and its transmission through tradition, experience and forms of apprenticeship. Gelwick (1996) in an overview of the life and work of Polanyi observes that:

> “Apprenticeship is a central example in the philosophy of Polanyi for showing that knowing is a personal activity with tacit coefficients …. Professional training in a community of experts who teach through their example and demonstrations was one of the clues to how that knowledge of "things we cannot tell" explicitly is passed on. There is an ocean of tacit coefficients that support the articulate parts of our knowing, and Polanyi had learned this in his medical studies.” (web page)

In doing so, he provides us with a summary of Polyani’s insights that are applicable to both connoisseurship and communities of practice. It is this link between connoisseurship and the community of practice in which it is situated, which provides judgements with both credibility and authority.

If either the community of professional practice or the wider community of practice to which this is related, does not accept that the connoisseur has demonstrated the expertise, authority and repeatability of judgement quality (primarily its comparability and consistency), then the judgements made will not be accepted as either dependable or credible. This means that it is the connoisseur who must meet the minimum standards for expertise and repeatability of judgement, rather than the task or the conditions for performance. This focus on the expertise and repeatability of judgement is important in any consideration of reliability but particularly in relation to assessments of performance. This is because although statistical measures of reliability may be rational and necessary for comparison between raters or different forms of assessment, they are unlikely to significantly alter either public perceptions of expertise and authority, or have meanings other than those that are socially determined. A Chief Examiner in a recognised public examination such as a General Certificate of Education (GCE) Advanced Level has by virtue of office, a mantle of authority and a perceived level of expertise and independence, that makes her or his judgements appear more credible than those of a teacher assessing coursework. Once again, we are reminded of Cronbachs' dictum that it is not the test but the inferences based upon it that are validated, only in the case of assessment by connoisseurship, it is the not the test but inferences arising from the judgements of the connoisseur that must be validated. This is because assessment by connoisseurship is not possible, unless there is a shared understanding of purpose and the authority and expertise of the assessor has been demonstrated and accepted beyond any reasonable doubt, by the community of practice and others involved.

Third, it means that the repeatability and relevance of the assessor's judgements, both on different occasions and over time must be maintained, if public confidence in the shared purposes of the assessment and in the judgements made by examiners is not to be reduced and the credibility of the examination affected.

In a connoisseurship model of assessment, an assessor is 'given permission to sit alongside' and make judgements. It is the nature of this consensual agreement, which characterises this form of assessment and distinguishes it from inspection or magisterial examination and judgement, both of which are externally imposed. Purpose gives shape to form, and in this case, to an examination in which proficiency, knowledge or ability is revealed for detailed inspection in order that it may be assessed and judgements made about its quality, value or appropriateness, in a process that 'piles up' the facts or evidence to see if the pointer or 'examen' moves. The slightest movement 'tips the balance' and this is another important characteristic of a connoisseurship model of assessment. First, because it allows the concept of a threshold, to function as a result of a judgement as to what is 'good enough', rather than a rule that is indiscriminate in its application. This is in accordance with Polanyi's (1962) observations first that:

> "Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art and cannot themselves either replace or establish that appreciation" (p31).

Secondly Polyanyi (1962) that:

> "Analysis may bring subsidiary knowledge into focus and formulate it as a maxim ... but such specification is in general not exhaustive. Although the expert diagnostician, taxonomist and cotton-classer can indicate their clues and formulate their maxims, they know many more things that they can tell…" (p.88)

A maxim is a rule that requires some knowledge of the domain in order to be understood and applied. Maxims are not strict rules because they require judgement in their application. This does not limit their value to an expert or connoisseur but it does mean that they are inadequate for the novice, who lacking the knowledge and expertise to use them, consequently requires strict rules, closely defined criteria or tight specifications to attempt a judgement modelled on that of the expert. To the novice the judgement of the expert appears to be an immediate, intuitive response, as Aristotle says in Physics, Bk. II, Ch. 8, "Art (*techné*) does not deliberate". Consequently, in a connoisseurship model of assessment, training to apply a test or to use criteria may be helpful but is not sufficient on its own as there must also be a level of mastery resulting from a form of apprenticeship as well as recognised expertise in the field being assessed. Indeed criteria may actually be a distraction for the assessor from the business of making a judgement about performance. This is because of the need to interpret criteria that are not necessarily a good fit with the nature of the performance being assessed, either because they are too tightly specified or are so general as to be more of a nuisance than an appropriate guide. Gipps and Stobart (1996) conclude that:

> “The main problem is that, as the requirements become more abstract and demanding, so the task of defining the performance clearly becomes more complex and unreliable. Thus while criterion-referencing may be ideal for simply defined competencies ('can swim 50 metres'), it is less so as the task becomes more complex: either the assessment must become more complex (for example, the driving test requires intensive one-to-one assessment) or the criteria must become more general. If the criteria are more general they are less reliable, since differences in interpretation are bound to occur.” (webpage)

Difficulties with criterion referencing are widely recognised and the monograph by Glass (1997) provides a valuable overview and discussion of these. One response has been the promotion of construct referenced assessment, Wiliam (1998) as a more appropriate basis for judgement. This may be an improvement on criterion referencing described by Wiliam (2000) and may well mitigate the effects of over specification in criteria as well, but only because it provides the assessor with the latitude to make judgements rather than to be misdirected by rules. However, this is not in itself sufficient to ensure the credibility of construct referenced assessment. To paraphrase Wiliam by replacing his word ‘criterion’ with ‘construct:

> “…in any particular usage, a construct must be interpreted with respect to a target population and this interpretation relies on the exercise of judgement that is beyond the construct itself. In particular, it is a fundamental error to imagine that what is described by the construct will be interpreted by novices, in the same way as it is interpreted by experts.” (webpage)

In the original, Wiliam is discussing the interpretation of a criterion by students in relation to statements by teachers intended to define behavioural competency, the ability of the student to respond being limited both by their own knowledge and experience and because criteria have no objective meaning independent of the context in which they are used. This is important because construct referenced assessment implies a connoisseurship model of assessment. Wiliam’s describes the practice of teachers involved in ‘high-stakes’ assessment of English Language for the national school-leaving examination in England and Wales and of the training involved as follows:

> “In this innovative system, students developed portfolios of their work which were assessed by their teachers. In order to safeguard standards, teachers were trained to use the appropriate standards for marking by the use of ‘agreement

> trials’. Typically, a teacher is given a piece of work to assess and when she has made an assessment, feedback is given by an ‘expert’ as to whether the assessment agrees with the expert assessment. The process of marking different pieces of work continues until the teacher demonstrates that she has converged on the correct marking standard, at which point she is ‘accredited’ as an assessor for some fixed period of time.” (webpage)

The parallels with a connoisseurship model of assessment are clear. Consequently, for this model of assessment to work, it is not enough to use construct referencing instead of criterion referencing but to recognise that in a connoisseurship model of assessment, the novice to mastery continuum applies to assessors, just as much as it does to students. This is because without the experience of apprenticeship implicit in connoisseurship, maxims cannot be interpreted and applied, first because the knowledge and expertise to use them is lacking and second because the requirements of purpose and authority, essential to a connoisseur and described previously are not met either.

This view that a form of apprenticeship and a level of acknowledged subject expertise is necessary finds support in the paper by Ecclestone (1999) who cites research from Nottingham University into the reliability of assessments in National Vocational Qualifications and concludes that:

> "… the variability of assessment judgements is caused by different interpretations of standards held by assessors."

Subsequently, reporting on research by others (Wolf 1995, Winter and Maisch 1996, Ecclestone 1996a, 1996b) she concludes that:

> "Social factors seem particularly important in developing teacher's ability to internalize a standard … that induction into an 'assessment community' helps them learn and internalize a notion of the right standard … and the need for informal, on-going dialogue about how standards should be interpreted within a community of subject based assessors." (p.57).

What is meant by the term connoisseurship when applied to assessment can be summarised as a form of assessment characterised by:

- assessment by a qualified person who is a member of a community of practice and whose authority as an expert in their field and as a connoisseur is recognised both within and outside of that community;
- the exercise by a connoisseur of critical faculties based on knowledge both within their field of expertise and as an assessor, that has been acquired, at least in part, by forms of apprenticeship;
- comparisons made in relation to perceived qualities in the work or performance being assessed, rather than comparisons made in

relation to other candidates or externally imposed standards or norms;

- purposes for the assessment that are shared and agreed both within and outside of a community of practice;
- the demonstration by the assessor of their expertise and authority as a connoisseur through the repeatability and relevance of their judgements on different occasions and over time;
- the exercise of judgement to determine what is sufficient for the award being considered to be granted and the candidate inaugurated into the community of practice that the award signifies.

The extent to which these six characteristics are met by a connoisseurship model of assessment and the means used to deliver it, seem likely to determine the dependability (meaning both validity and reliability) of the assessment and the credibility of the award(s) derived from it.

For a connoisseurship model of assessment to function, there must be agreement individually and collectively in the community of examiners, not only of what it means to be competent in a particular domain but also that this must be exemplified, in order for it to be shared and applied consistently and comparably. This has far reaching consequences, as it affects not only the way assessments are made (e.g. choice of criteria, statements, standards, weightings, rubrics) but also the meaning of reliability and the ways used to quantify it.

What might reliability mean in relation to a connoisseurship model of assessment?

Statistical methods of estimating reliability based on correlations are not appropriate for a connoisseurship model of assessment. The fact that numerical scores make quantitative methods for evaluating the end result available does not mean that these methods are always appropriate or that difficulties with applying and interpreting the results do not exist. One reason for this is that the use of numbers can provide notions of ‘accuracy’ and ‘measurement’ that only tell part of the story. Where numbers are averaged or aggregated this problem with ‘accuracy’ becomes more acute as decision consistency is made more complex and the reasons for decisions are made less accessible. Another reason is the reliability of the judgements that provide information for the assessment. Difficulties arise because of the complexity of the judgements that have to be made, the extent to which inferences are drawn and the assumptions on which these are based. For instance it is not unusual for discussions about the quality and consistency of assessment decisions to be conducted in terms of sufficiency of evidence, its diversity and relevance, the range of contexts in which it has been produced and the assessment methods used. These methods may involve observation, questioning candidates, judging products, evaluating records and taking into account information from self-assessment items. The process of assessment may involve some or all of these methods together with decisions about sufficiency and appropriateness of evidence and professional judgements about factors particular to each candidate. As the use of less precise criteria increases, more and more sources of variation are introduced into the assessment. This is the world of construct referenced assessment and expert judgement and in these circumstances the meaning of reliability depends on context, performance variables and the quality of assessor decisions.

Difficulties with the meaning of reliability in relation to assessments of performance also arise because assumptions and inferences drawn from the terminologies in common use to describe assessment methods and reliability, are subject to change over time and across cultures. These changes are related to philosophical and cultural understandings and result in differing concepts of the meanings and relative importance of factors like validity and reliability. For example, in a description of change relating to criterion referenced assessments Griffin(1998) refers to the corruption of criterion referenced approaches to assessment that had arisen in the 1980s in comparison to the way it was intended to be applied and understood in the 1960's. This corruption had given rise to an atomistic, if 'it can't be stated it doesn't exist' attitude summed up by Jessup(1989) as:

> "… reliability and validity have become a cliché… are separate although related concepts … their significance is quite different in the standards-based, criterion-referenced model of assessment which applies to NVQ's … we should just forget reliability altogether and concentrate on validity, which is ultimately all that matters." (p. 191);

Jessup(1989) subsequently concluded that reliability is yet another part of the baggage people carry with them from traditional norm-referenced models of assessment. This was a view that found ready acceptance at a time when political imperatives meant that a greater involvement of business in education was sought, together with increasing levels of control and standardisation. In less than thirty years an approach to assessment, clearly rooted in classic psychometrics (with all that is implied in this for traditional views of reliability) that had set out to make scores informative about behaviour rather than merely about relative performance, had migrated to become a standards referenced methodology in which reliability was just 'baggage'. The comment from Glass (1977) sums up what has happened:

> "The evolution of the meaning of "criterion" in criterion-referenced tests is, in fact, a case study in confusion and corruption of meaning." (webpage)

Understanding the relationship between a connoisseurship model of assessment and other forms of assessment is important to an understanding of what reliability means in this context and to the development of a means of quantifying reliability that is appropriate and credible.

If the types of assessment in general use are considered to lie on a continuum between ‘pure’ criterion referencing and expert judgement or connoisseurship, then the consequences for the way the method is applied and the extent to which a judgement may be exercised by an examiner, rater or assessor, can be visualised in the format shown in Figure 1. Visualising the basis for assessments in this way is a reminder that the tendency to describe and think of them as distinct types or methods is not correct or helpful in any consideration of reliability, as they are all in effect, fuzzy sets. Moreover, because:

i. connoisseurship may employ in varying degrees, both constructs as references, and criteria for definitions (even if these are tacitly understood rather than explicitly stated);

and

ii. criterion referencing may (especially in less specified forms), require both reference to domains and the expert, critical judgements that are a hallmark of connoisseurship;

then it may be concluded that in the process of assessment, there is no such thing as the application of either criterion referencing, construct referencing, or connoisseurship, as distinct and separate kinds of procedures but that they are all parts of a larger fuzzy set.

As a point of reference, assessments of coursework and of essay type responses in (for example), GCSE examinations probably fit within the construct referenced ‘zone’ but even within the same examination, different components may be either more or less construct referenced depending on the techniques employed to record the judgements required.

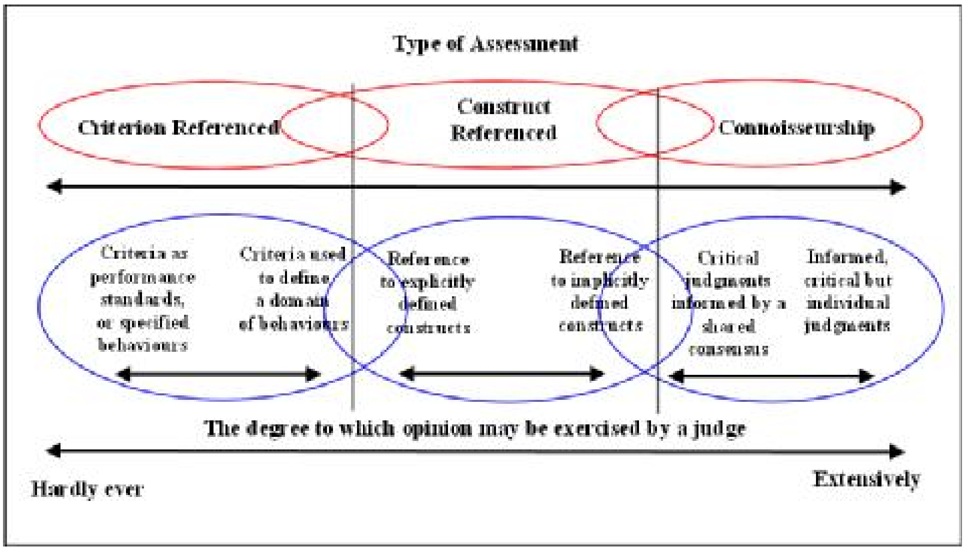
Type of Assessment
Criterion Referenced
Construct Referenced
Connoisseurship
Criteria as performance standards, or specified behaviours
Criteria used to define a domain of behaviours
Reference to explicitly defined constructs
Reference to implicitly defined constructs
Critical judgments informed by a shared consensus
Informed, critical but individual judgments
The degree to which opinion may be exercised by a judge
Hardly ever
Extensively

Figure 1. A continuum - Criterion referencing to connoisseurship.

Even if it were possible to regard each type as being in some way distinctive, then as the diagram indicates, the way assessors make judgements in each type of assessment also exists on a continuum. For instance, criterion referenced assessments range from tightly specified pass/fail criteria related to vocational competencies, or criteria used as cut-off scores, to descriptions along a continuum of achievement. Any attempt to measure reliability needs to take account of this complexity.

Moreover it is not enough to assert that the issue of reliability can simply be disposed of by a dependence on the evidence of performance Jessup(1989), if only because sufficiency of evidence requires expert judgement by an assessor. Inferences are made by people, so if the original judgements and any subsequent inferences are not consistent and comparable, then the results of the assessment are unreliable. This example of change in the meaning and application of criterion referencing over time suggests that the changes owe more to philosophical, political and social factors than considered development or systematic research and application and that, the consequences of these changes have far reaching implications that are not confined to philosophical and political dimensions but have a direct influence on the practice of assessment. Of more immediate relevance is the fact that changes in the meaning of a type of assessment serve to increase uncertainty about the way it applied by those making an assessment. Griffin (1998) concludes that:

> "Very few teachers can articulate the meaning of the assessment approach required by criterion referencing. Most still define it with the approach used in the 1970s." (webpage)

So not only does the nature of the judgements made using a type of assessment range along a continuum within the type but the way the type of assessment is being used may vary from assessor to assessor, even though there is an assumption that it is applied in comparable ways by all involved. The comment by Gipps (1994) that:

> " … to achieve justice in assessment, interpretations of student's performance should be set in the explicit context of what is or is not being valued on the basis of what evidence or prejudice." (p.265);

is a reminder of the wider consequences both of the lack of certainty about what is being done, to whom and why, and because what is accepted as being 'reliable' also depends on values and judgements. Broadfoot (1999) describes assessment as:

> "Essentially... a 'technical craft' but …. a social technology (Madaus, 1994) and in that sense, it is not the techniques themselves that need to be a focus for concern, but how they are used" (p.3).

Broadfoot(1998) discussing quality standards and control in higher education observes that assessment, is not, and cannot be, simply the application of a neutral technology because assessments are not valid in themselves, objective or independent but interact with what they are supposed to measure, Torrance(1994) makes similar observations in a more general context.

What is being illustrated is just how complex the business of assessing performance actually is and as a consequence why the apparently objective measures of reliability that are frequently published should be treated with caution. Levels of statistical significance and measures of correlation in regard to assessments of performance are suggestive of a degree of validity that does not exist in reality. To paraphrase the comment by Gipps (1994) cited previously, measures of reliability should be set in the explicit context of what is or is not being valued and on the basis of what evidence or prejudice is known to be present.

How might reliability be expressed in relation to a connoisseurship model of assessment?

There is an extensive literature relating to concepts and measures of reliability both within and outside of traditions in education and the social sciences. For example, in industrial statistics, reliability denotes a function describing the probability of failure as a function of time; in process engineering it is the repeatability and reproducibility of the processes and the extent to which variability can be kept within bounds. Human factors engineering provides examples that combine aspects of both of these and approaches related to psychometric theory and their application in aviation and other high risk activities where human judgement about performance is required. During the course of this research more than 465 documents relating to reliability, mostly published on the worldwide web but also in books and journals have been reviewed in order to both identify if established means of expressing reliability relevant to a connoisseurship model of assessment exist and to develop a model to express the reliability of assessments of performance in graded examinations.

Wiliam (1997) proposes the use of Receiver Operating Characteristics (ROC) analysis as a means of measuring the decision consistency of an assessment system noting that:

> “The important point about the ROC curve is that it provides an estimate of the performance of the assessment system across the range of possible settings of the ‘standard of proof’. The ROC curve can then be used to find the optimum setting of the criterion by taking into account the prior probabilities of the prevailing conditions and the relative costs of correct and incorrect attributions.” (webpage).

The underlying assumption is that the assessor acts as a ‘receiver’ of signals and that the proportion of these that are misclassified is an indication of system reliability as well as a means of standard setting. Murphy (1982) in a report on the reliability of marking in GCE examinations notes the desirability of alternative measures to the correlation coefficient and cites McVey (1976) as suggesting the use of the signal to noise ratio for this. This is an earlier name for the ROC technique and Murphy (1982) concludes that this had not received wide spread acceptance. The reasons for this are not stated but in relation to an assessment of performance, there are appear to be three possible objections to its use.

First, that it assumes that unreliability is due to variations in assessor judgements and that what remains is ‘noise’. This discounts sources of unreliability that are due to other factors of which faulty ‘equipment’, for example the nature of the task, rubric, specifications or quality of the marking scheme are just a few. These are not necessarily matters of validity because they may well be valid, yet still be a source of unreliability because of their operating characteristics. These and related factors are all amenable to investigation and control in varying degrees and for the extent of reliability to be stated in any meaningful way this is necessary as for example, it is not impossible for assessors to be using an unreliable ‘tool’ in a consistently reliable manner.

Second, that it assumes that the assessor is a ‘receiver of signals’ and that there is no interaction between transmitter and receiver. This may be the case in for example in the marking of a written script from a history examination (although even this is doubtful in one sense as research into the influence of gender or handwriting suggests), but it certainly not the case for assessments of performance in music or dance. This is particularly true of most graded examinations, where a level of interaction from the assessor as audience for the performance, is a necessary part of the validity of the examination.

Third, and of less importance, that its appropriateness as a standards setting mechanism is open to question if only because of the difficulties raised by the assumptions noted above.

The work on generalisability theory by Cronbach, Glaser, Nanda, and Rajaratnam (1972) is well known and widely cited. Cronbach, Linn, Brennan, and Haertel, (1995) provide a valuable overview of the application of generalisability analysis to educational assessment and indicate some of its limitations. Generalisability analysis allows characterisation of the variables that affect the reliability of an assessment, based on a statistical theory describing how multiple sources of error in measurement can be estimated separately in one analysis and permits predictions for the accuracy of similar tests with different numbers and configurations of assessors, components, constructs and similar factors that comprises an assessment. Separating out different potential sources of variation provides indications of how these different sources of variance arise and contribute to an understanding of how various aspects of the assessment could be improved. As a consequence, it provides measures of which type of assessment, or components of an assessment result in more reliable scores and can be used to examine aspects of inter-rater reliability. Its limitations with regard to assessments of performance that emphasise construct referenced judgements by examiners arise largely because of the assumptions that it requires in relation to scores, decision rules, and disagreement amongst assessors. Other developments such as Item Response Theory may be used to address this problem but the fundamental problems with assumptions in relation to assessments rather than ‘tests’ remain. There is a sense in which it provides the right answers to the wrong questions and this is because of its roots in the American tradition of measurement rather than the European one of assessment. A consequence of this psychometric foundation is the shift the focus from quality of judgements required in the European tradition, to the reliability of inferences drawn from the task.

What is understood as reliability and the consequences of those understandings are not just affected by social factors or the nature of an application but much more profoundly by the way in which the whole question of meaning in relation to reliability is conceptualised. So as an analytical tool, generalisability analysis is clearly valuable and of wide application but because of the necessity for judgements in relation to construct referenced assessment, it is likely to be of limited value in generating measures of reliability, partly because of the underlying assumptions and necessary conceptual differences and partly because it is not easily used or understood by assessors.

Fourali(1997) proposes an alternative approach that attempts to take into account the uncertainties inherent in the application of construct referencing to assessments of performance. Drawing on understandings from the use of fuzzy logic in describing how well a value judgement conforms to a semantic ideal, it recognises that the decisions made by an assessor can never be present/not present but always incorporate a degree of uncertainty. The context for this is the application of criterion-referenced decisions to portfolio assessment in vocational qualifications. The criteria used are sufficiently generalised (e.g. relevance, sufficiency, variety) to indicate that they require a level of judgement that implies the use of underlying constructs by the assessor or that they are used to define a construct that is being applied as part of the assessment process. A method of calculating a standard deviation that takes into account the uncertainty or fuzziness (SD fuzzy) is described and the conclusion drawn that the smaller this measure is, the greater the certainty of the assessors' rating. The objection to this is that once again it makes the assumption that unreliability is largely due to variations in assessor judgements. In doing so it raises the same factors of the nature of the task, rubric, specifications or quality of the marking scheme and does not take into account the range of interactions that exist between them. These are in themselves a source of uncertainty or fuzziness, both for the assessors and for any inferences as to how consistent and comparable the assessments actually are. That said it might be that reporting an SD (fuzzy) index would provide an indication of the degree of uncertainty in judgements based on construct referencing.

Two areas provided examples that appear to be both relevant and applicable to concepts of reliability in relation to a connoisseurship model of assessment. The first is research into the assessment of performance by airline cockpit crews in training, the second is applications of Statistical Process Control as developed and practiced in precision engineering.

Johnson and Goldsmith (1998) and Holt, Johnson and Goldsmith (1998) describe methods of assessment in relation to assessments of aircrew performance. There are some similarities between the sort of assessment described by these authors and those that take place in an educational context and more particularly, to those assessments of performance in music and dance where observation by an assessor is common practice. Assessing the performance of aircrew, for example in relation to safety and the management of resources during a flight, is clearly a matter of more consequence or ‘higher stakes’ than assessments of performance in more general educational settings, so it is reasonable to expect that reliability of assessments is of equal importance. What follows is a brief outline of the context for the assessments described in order that the similarities between this and a construct referenced approach to assessment may be considered.

Aircrew training makes extensive use of flight simulators that enable a range of actions, circumstances and situations to be simulated realistically in the aircraft used and the responses of aircrew to these assessed. Assessments of performance are undertaken by evaluators (or to use the term in a European way, ‘assessors’), who make judgements against agreed standards and record their decisions. Standards are set in relation to the standard operating procedures of the airline, which may match industry wide standards and in any case will have been approved by the regulatory authorities, for example the Federal Aviation Administration (FAA) in the United States. These standards are carefully specified and provide the same basis for training, operational practice and assessment. This process of training and assessment is described in an FAA review of research (Edens 1997) projects as:

> “Line-Oriented Evaluations (LOEs) are a methodology used in Advanced Qualification Programs (AQPs) to evaluate pilot training performance and establish trainee proficiency. LOEs consist of flight simulation scenarios that are developed by the training organization and approved by the FAA.” (p.19).

And by the FAA (1990) as:

> “Line Operational Evaluation is primarily designed for crewmember evaluation under an Advanced Qualification Program (AQP). Line Operational Evaluation is conducted in a flight simulator or flight training device and is designed to check for both individual and crew competence. Line Operational Evaluation may also be used to evaluate a specific training objective.” (webpage)

A paper by the Human Factors Group of the Royal Aeronautical Society (1990) describes the basis for this type of training and assessment as follows:

> "…crew performance will be determined by individuals behaving and operating to a set of standards; which will require them to have certain knowledge, skills and attitudes. Developing this knowledge, these skills and attitudes in individuals, will be dependent on trainers behaving and operating to certain standards, and likewise this will require them to have the commensurate knowledge, skills and attitudes." (webpage).

The qualities to be assessed are defined as:

Knowledge	(e.g. self, roles, systems)
Skills	(listed as: Communications, Effective Teamwork, Task Management)
Attitudes	(defined as: Values and beliefs which influence people to select a set behaviour)

These are assessed by reference to 'Crew Competence Standards' that are also referred to as Standards of Competence and Behavioural Markers. Selected extracts from typical behavioural markers / competency standards that various organisations are currently using are provided in order to illustrate what these mean and the levels of competence implied. Extracts taken from these are indicative of what is to be assessed:

1. “ British Airways behavioural markers:

- Tone of flight deck is friendly, relaxed, supportive.
- Crews adapt to other members personalities.
- Crews act decisively when situation requires.”

2. “NASA /UT LOS Checklist

- When conflicts arise, the crew remain focused on the problem or situation at hand.
- Crew members listen actively to ideas and opinions and admit mistakes when wrong, conflict issues are identified and resolved.
- Crew members verbalize and acknowledge entries to automated systems parameters.
- Cabin crew are included as part of team in briefings, as appropriate, and guidelines are established for coordination between flight deck and cabin.”

3. “Management Charter Initiative Level 4 Competency Standards:

- Individuals are encouraged to offer ideas and views and due recognition of these is given.
- Information about problems is clear, accurate and provided with an appropriate degree of urgency.
- Potential and actual conflicts are identified and actions promptly taken to deal with them.
- Inadequacies in information are identified and alternative sources are sought.” (webpage)

Similar competencies are given from other airlines and from the CRM NVQ Level 4 Competency standards. This information has been given in some detail in order to draw out the similarities between assessments in ‘Line-Orientated Evaluation’ and the nature of assessments undertaken in an educational context using construct referencing.

In practice, LOE assessments appear to be employing domain referenced behaviours but the distinction between this and construct referencing is unclear. A construct can be defined as a hypothesised trait, ability or characteristic that is abstracted from different but observable behaviours that act as signs, proxies or indicators of what is being conceptualised. This does not seem significantly different to a well defined behaviour. Both appear to be hypothesised, denotative and open to interpretation, a fact acknowledged by the extent of the literature on both construct validity and on definitions of behaviours. It seems probable that it is only the assessment context and the way judgements are recorded that is likely to distinguish them, with domain referencing lying towards the criterion referenced end of the construct referencing continuum rather than towards a more generalised form of assessment by connoisseurship. If this is accepted, then it is not unreasonable to infer that a method of quantifying reliability used for aircrew LOE and training may be appropriate for assessments of performance in an educational context.

The means of doing this proposed by Johnson and Goldsmith (1998) is that of a 'multi-pronged' approach where the meaning of reliability as a measure of consistency, is extended to include properties of sensitivity and accuracy. These are defined as:

> "Sensitivity refers specifically to the degree to which observations track or covary with changes in the object being measured" (p.2)

> "Accuracy refers to how closely our measurements correspond to the absolute magnitude of what is being measured" (p. 3)

Two methods are presented for assessing the reliability of observations, rater-referent reliability (RRR) and inter-rater reliability (IRR). The authors state that:

> "Although both methods can be said to measure reliability, we believe RRR is the better measure of sensitivity. Also, the reader should be forewarned that these labels (RRR and IRR) are somewhat misleading in that they suggest they are measures of rater reliability, when in fact they also reflect the influence of the measuring instrument and various other factors that influence the sensitivity of the observations." (p.7)

Rater-referent reliability is described as a correlation reflecting how closely an evaluator's ratings agree with some standard or referent to be used when there is an external, objective basis for defining a referent score. Inter-rater reliability as a correlation reflecting the degree to which a group of raters agree with one another and as the most commonly used method of measuring rater reliability and one that does not require a referent value. The authors propose that the two measures be applied together with rater-referent reliability as the primary measure of sensitivity and inter-rater reliability as a means of diagnosing rater-referent reliability that is lower than expected. Recognising the subjective nature of an assessment process the authors conclude that a process of training and calibration will minimise this, they also note that group of assessors might deviate from the referent for valid reasons and recommend checking for deviation between the referent and the group's averaged ratings.

Additionally they propose the use of the mean absolute difference to estimate the accuracy of the observations as this may be used in situations that do not easily lend themselves to correlational analysis. A frequency analysis of the degree to which assessors are using the rating scale in a manner that is congruent to the referent is also suggested as a diagnostic tool when mean absolute difference and rater-referent reliability is lower than expected. The description of how these measures are derived (pages 7 – 14), indicates that the first stage is to create a grading sheet listing the indicators prescribed by the crew competence standards together with a Likert style scoring scale. This then used in an agreement trial by a group of experts (training supervisors) viewing a video-tape of the performance to be assessed in order to provide a referent value for each performance indicator. Next the same video tape is viewed by a second group of experts (the assessors) who use the grading sheet and ratings scales to rate the same performance indicators. Inter-rater reliabilities are then computed for the assessors and an overall inter-rater reliability figure is derived from this. Correlations between each assessors ratings and the referent score is calculated and the average of these used as a measure of the overall referent-reliability of the assessment process. This then becomes a standard and deviation from it a measure of the reliability of the assessment. The process of deriving a referent score has similarities to those used in the GCSE examination for the visual and performing arts for the development of marking schemes although the ways that these are used and are subsequently ‘standardised’ are different.

The use of a referent in this way requires an agreed set of criteria or indicators and an agreed overall standard for the activity to be assessed, either as part of the performance or as the performance itself. It also presupposes that assessments occur on a regular basis and not on one or two occasions a year. This probably limits its application in regard to assessments in large scale public examinations, although the increasing use of common standards and specifications for these such as those being developed by the Qualifications and Curriculum Authority, in England and Wales may this less of a problem. Within specific subjects such as Mathematics and Science it may be easier to use and for task specific requirements such as those found in vocational education it would appear to have advantages. What this approach to reliability does is to offer a methodology and measure that may be of practical value for assessments of performance but further research into this would is clearly necessary and is well outside the scope of that being reported on here. It also indicates that deviation from a referent, or ‘absolute standard’ (sometimes referred to as a ‘gold standard’) is practicable as a means of conceptualizing reliability in relation to construct referenced assessment. This could be objected to on the grounds that it is not possible to create an absolute standard, either for a construct or for a performance and that consequently its use is impossible. However, if the assessors were able to generate this, either in the way previously outlined or as a result of the process of standard setting by teacher assessors described by Wiliam (2000) then this may be feasible. An alternative approach to this is described subsequently but before coming to this, the application of Statistical Process Control to reliability in the context of construct referenced assessment is considered.

Statistical process control (SPC) is a simple, yet powerful, collection of tools for graphically analysing process data that was invented in the late 1920's by Walter Shewhart in order to monitor and improve processes. Originally intended for use in a manufacturing environment it was subsequently extended by W. Edward Deming to improvement in all areas of an organization. Outside of education, it is well known and widely used, especially in engineering although its applications are by no means limited to that. For example, its use in medicine is described by one consultancy group as:

> “SPC represents a shift in the way we think about measuring performance and analyzing data. The traditional approach, strongly emphasized in clinical research, collects data and then compares the data to either some past data set or a control group data set. While this is perfectly legitimate in clinical research trials, it is insufficient for measuring and improving performance in real time. Real time data collection and analysis means measuring, tracking, and assessing performance everyday. SPC provides the simplest and most powerful tools for real time performance assessment” (webpage)

Statistical process control uses statistical analysis of individual process measurements to categorise performance variations in one of four ways: common cause, special cause, compensation and structural. Common cause variations are the result of everyday, uncontrollable influences, special cause variations are sporadically occurring factors that send performance outside the range of common cause variation, compensation is a cause of variation arising from attempts at control and structural variations occur systematically because of cycles or trends. The emphasis of statistical process control is on the reduction or elimination of special cause factors and inadequately controlled compensations, so it provides a means of more accurately measuring performance against flexible standards and by defining common cause variation and the limits of this, establishes acceptable performance ranges. Statistical process control makes extensive use of the graphical representation of quantitative information. Typically, this information is provided as one or more process control charts used to plot a function of process measurements against time. Points that are plotted on the graph are compared to a pair of control limits in order that the process may be both monitored and improved, control charts represent a compromise between the risks of not detecting real changes and of false alarms and for that reason the choice control limits needs careful consideration.

Some evidence on the use of statistical process control in relation to management practices in educational settings is available. One example is a report by Shor and Robson(2000) on an investigation into a continuous improvement process based on feedback control for examining individual performance of a student and modifying student's learning experiences. However, published literature relating to the application of statistical process control methods to examinations or assessment appears sparse. Konrad (1998) in a wide-ranging and useful review of issues relating to the delivery and assessment of vocational education in the United Kingdom concludes that:

> "The purpose of sampling assessment decisions is to provide evidence of the consistent application of the assessment and verification process. In large Assessment Centres, it is theoretically possible to use statistical techniques to guarantee consistency. However, such an application of a Statistical Process Control model is not only unlikely to be valid given the range of circumstances, but more seriously, assumes that the process outputs are sufficiently capable of exact measurement." (webpage)

This inference drawn from this author's conclusions is that assessors, verifiers and others have difficulties in understanding and applying the assessment and verification processes adequately. Given the information provided this may well be the case but the inference that a statistical process control would not be valid needs to be treated with caution. First, because the author is arguing for an alternative approach based on the Quality Assurance based approach of continuous improvement exemplified by the European Quality Foundation model for Total Quality Management (EFQM) model, as opposed to the process-control model of ISO 9000 which is dismissed as incompatible with the mission and goals of many educational organisations. Second and more significantly, because no evidence or analysis is provided to support the inference made. Even, as seems to be the case from the contexts illustrated in the paper, the locus for monitoring is the educational organisation rather than the awarding body. The assumption that 'the process outputs are sufficiently capable of exact measurement' appears to misunderstand the purposes of statistical process control. For instance, Kerridge and Kerridge(undated) describe the application of statistical process control as:

> "The aim is first of all to find out if the process is stable. If it is stable, that is, "under statistical control", the aim is to set priorities for investigation. The investigations are to find ways of changing the process, by permanently removing "special causes" of variation. If, on the other hand the process is already stable, we can use the control chart to demonstrate the effect of experimental changes." (webpage).

Their view of statistical process control presents two subtle but important shifts in emphasis, one from 'control' to 'process' and the other from 'outputs' to 'inputs'. This is of significance to any consideration of ways to apply statistical process control in the context of assessment, not least because it enables the purpose to be shifted from a narrow concern with ‘outputs’ to a more proper consideration of how the process of assessment is to be managed. This view that assessment is a process and not an event is fundamental to this.

The steps in the production of the statistical process control charts that are the basis of the information provided are well documented, examples of this are the Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook) and statistical analysis programmes such as Minitab software (Release 13 2000 – www.minitab.com), which provides extensive and well documented facilities for statistical process control. Consequently, only a brief summary is provided here in order that the basic requirements can be related to the context of construct referenced assessment. The first requirement is to identify the process parameter that is to be monitored (such as the process mean, or spread), this is then used as the centre line of a plot set according to the target value required for the parameter. The second requirement is to group representative measurements as sets (in industry by time period) and to then plot these points on a chart and relate these to the process parameter. For example if the mean scores (representative) of an assessor in relation to a mean score of the universe of assessors (parameter) is of interest over a sequence of occasions then the plot is of the sample-means, computed for each occasion. If the point to be plotted is an occasion O, denote this as Xo, then, create upper and lower control limits (UCL, LCL) in accordance with the formula, UCL = CL + 3 sd, LCL = CL - 3 sd, where sd is the standard deviation of Xo. In this example, Xo is the mean of all the final scores awarded by the assessor during a single examination session. If each occasion sample comprises of n measurements, then the standard deviation of Xo is equal to the process standard deviation divided by the root of n. After the control limits have been established a sequence of points (occasions) continue to be plotted for the assessor. When a point goes outside of the control limits, it indicates that there are factors that need to be investigated. For example it may be that it is a false alarm (the British Standard uses "3.09 sigma" limits (corresponding to 2% of false alarms), or that the centre where the examinations take place has significantly higher or lower results for these examinations than the universal mean, or that there were other external factors at work that should be investigated. If no other factors were at work and it is not a false alarm because of administrative error, then it means that the decisions of the assessor should be reviewed, especially if there was evidence of random error or an underlying cumulative trend. The same charts can be plotted for all examiners and inspected to indicate deviation from the parameter, as well as for subjects, centres, schools or the awarding body depending on the choice of process parameter. Note that the purpose of the use of statistical process control in this context is management rather than ‘control’.

Conclusions

If the methods and concepts of Rater Referent Reliability described previously and methods and concepts drawn from Statistical Process Control are synthesised , two possibilities emerge. The first is that an alternative meaning for the term reliability in relation to the European tradition of construct referenced assessment may be proposed as being:

The stability of rater judgements relative to a referent defined as the periodically reviewed universal mean score derived from the mean final scores awarded by a representative sample of raters in each setting (e.g. single subject, examination, grade or level, group of subjects, grades or examinations).

This makes the stability of rater judgements within bounds set by the community of practice, the determining factor in measures used to express the reliability of assessments of performance. In the case of examinations, centres, schools or awarding bodies the phrase ‘rater judgements’, would be replaced by the object of interest.

Stability means that over time and on each occasion, the results for both the candidates and for the examination remain within these bounds. Gipps (1994) discussing inter-assessor reliability and test-retest reliability described this as:

> "The extent to which an assessment would produce the same, or similar score if it was given by two different assessors, or given a second time to the same pupil using the same assessor." (p. 2)

In a construct referenced or connoisseurship models of assessment reliability the extent of 'sameness' may be described as the amount by which assessment decisions may be vary and still be regarded as consistent and comparable, rather than deviating to an extent that renders them unacceptably inconsistent. Gipps also states that:

> "…one outstanding problem which we have in assessment is how to reconceptualise traditional reliability (the 'accuracy' of a score) in terms of assuring quality, or warranting assessment based conclusions, when the type of assessment being used is not designed according to psychometric principles and for which highly standardised procedures are not appropriate" (p. 2).

The phrase "the extent to which" is crucial to this because if 'extent is not stated, then the quality of an assessment is open to question and results are not 'warrantable'. The second possibility arises from this and is the application of the concept of a region of acceptably stable results to the setting upper and lower control limits or 'bounds' on the extent to which the results of an examination may vary and still be accepted as reliable in the context and for the purposes of an examination. Setting bounds and demonstrating the stability of results provides a means of clearly stating what reliability means in a particular setting and of managing the process of assessment and standardisation to ensure that both processes and results can be shown to correspond with this. Incorporating management of the processes of assessment and the processes of standardisation into this definition of reliability, takes into account the system through which assessment judgements are made. This is important because whilst the proposed definition emphasises the stability of rater judgements, the stability of the processes in which judgements are embedded directly contributes to this. They are in effect two sides of the same coin because of a reciprocal relationship between the stability of rater judgements and the stability of the processes that enable the judgement to be made. For both types of stability to be maintained, systematic and adaptive strategies are required and these need information on the state of this reciprocal relationship as well as the factors that are at work in it. Expressed crudely, this is a feedback loop, however in practice it is considerably more complex than this and is probably better thought of as the sort of organisational learning described by Argyris and Schon (1974), and Argyris and Schon (1978) as well as the sort of processes and relationships described by de Geus (1997) or Senge (1990). This ‘feedback’ is used to optimise the stability of judgements, first by systematically monitoring and adjusting the technical and functional factors that contribute to optimisation and second, by improving the training of assessors.

This paper has described applications of statistical process control to a form of assessment that belongs to a predominately European tradition that is characterised by the use connoisseurship and expert criticism. This model of assessment results in judgements that are predominately estimative and indicative. Applications of this model result in assessments of performance rather than performance based assessments. The extent to which one or other of these assessment paradigms is being applied is determined first by purpose and second, by the extent to which opinion may be exercised by the person responsible for the judgement. In the first paradigm statements about reliability, relate to the extent to which the test is standardised and valid and the extent to which instruments and measures used have predictive value. In the second paradigm statements about reliability, relate to the extent to which an assessor makes judgements that are considered repeatable and credible by a community of practice. Inferences may be drawn from results derived by either tradition and scores used to rank order, grade or norm reference candidates with varying degrees of validity. Both paradigms occupy different ends of a continuum that extends from criterion-referenced measurement to assessment by connoisseurship. In the central area of this continuum, the paradigms may overlap as various forms of construct-referenced assessment are applied.

The credibility of classical measures of reliability arising from a psychometric tradition is diminished as assessments become more construct-referenced and indicative. Construct-referenced assessments of performance that place greater emphasis on the use of expert judgement require the use of alternative measures of reliability that are indicative of the repeatability of assessor judgement. Conceptualising or estimating reliability as the extent to which the stability of rater judgements and results relate to a referent and remain within bounds set by the community of practice is a useful approach in relation to construct-referenced assessment. This is because it focuses on the repeatability and reproducibility of both judgements and results and thereby corresponds with the view of reliability expressed by Gipps (1994) as:

> *"The extent to which an assessment would produce the same, or similar score if it was given by two different assessors, or given a second time to the same pupil using the same assessor." (p. 2)*

The concept of reliability as the stability of rater judgements and results allows techniques based on Statistical Process Control, Receiver Operating Characteristics and Generalisabilty Theory to be used, either singly or in combination in order to generate measures of reliability applicable to assessments of performance. It also permits unreliability to be conceptualised as a lack of stability and for this concept to form the basis for questions about sources of unreliability in assessments of performance.

References

Argyris, C. and Schön, D. (1974) Theory in practice: Increasing professional effectiveness, San Francisco: Jossey-Bass.

Argyris, C., & Schön, D. (1978) Organizational learning: A theory of action perspective, Reading, Mass: Addison Wesley.

Broadfoot, P., (1998) Quality standards and control in higher education: what price life-long learning? International Studies in Sociology of Education, Vol. 8, No. 2, 1998

Broadfoot, P, (1999) Empowerment or Performativity? English assessment policy in the late twentieth Century. Paper delivered as part of the Assessment Reform Group Symposium on Assessment Policy, British Educational Research Association Annual Conference University of Sussex at Brighton 1999 http://www.leeds.ac.uk/educol/documents/00001216.doc

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley: New York

Cronbach, L, J., Linn, R, L., Brennan, R, L., and Haertel, E., (1995) Generalisability Analysis for Educational Assessments. Evaluation Comment. UCLA's Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing. 1995 www.cse.ucla.edu/CRESST/Newsletters/CSU95.PDF

De Geus, A., (1997) The Living Company, Boston, HBR Press.

Ecclestone, K. (1999) 'Empowering or ensnaring?: the implications of outcome-based assessment in higher education' Higher Education Quarterly 53(1).

Edens, E., (1997) Air Carrier Training Research Review June 1997, FAA Office of the Chief Scientist for Human Factors (AAR?100). Project Title: Rapidly Reconfigurable Event-Set Based Line-Oriented Evaluations (LOE) Generator.

Eisner, E, W., (1998) The Role of Teachers in Assessment and Education Reform: Speech presented at the BCTF AGM, March 1998. http://www.bctf.ca/publications/speeches/eisner.html

Eisner, E. W., (1985) The Art of Educational Evaluation. A personal view, Barcombe: Falmer

Eisner, E. W., (1998b) The Enlightened Eye. Qualitative inquiry and the enhancement of educational practice, Upper Saddle River, NJ: Prentice Hall

Fourali, C. (1997). Using Fuzzy Logic in Educational Measurement. Evaluation and Research in Education, 11(3), 129-148.

Gelwick, R.,The Calling of Being Human, Polanyiana (Budapest), 5(1), 1996, pp.63-75. www.chemonet.hu/polanyi/9601/calling.html

Gipps, C., (1994) Beyond Testing Towards a theory of educational assessment. London: The Falmer Press

Gipps, C., Stobart G., (1996) Developments in GCSE. Journal of Educational Evaluation. Volume 4. 1996 http://www.aseesa-edu.co.za/newpage5.htm

Glass, G, V., (1977) Standards and Criteria Paper #10, Occasional Paper Series University of Colorado, December 1977.
http://www.wmich.edu/evalctr/pubs/ops/ops10.html

Griffin, P., (1998) Outcomes & Profiles: Changes in Teachers' Assessment Practices, Curriculum Perspectives, Vol 18, No. 1, 1998 pp 9-20

Holt., R., Johnson, P. J., & Goldsmith, T. E. (1998). Application of psychometrics to the calibration of air carrier evaluators. FAA Technical Report. www.faa.gov/AVR/afs/afs200/afs230/aqp/rhotlpap.pdf

Human Factors Group of the Royal Aeronautical Society (1990) QUALITY CREW RESOURCE MANAGEMENT http://83.146.40.74/crm/reports/qual-crm.htm

Jessup, G. (1991) Outcomes: NVQ’s and the emerging model of education and training. London: The Falmer Press.

Johnson, P, J., and Goldsmith, T, E., (1998). The importance of quality data in evaluating aircrew performance. FAA Technical Report. www.faa.gov/avr/afs/ratterel.pdf

Kerridge, D., and Kerridge, S., (undated) Two types of SPC. A working paper of the British Deming Association Statistics Group. http://deming.eng.clemson.edu/pub/den/files/2spcs.txt

Konrad, J., (1998) Assessment and Verification of NVQs: policy and practice. Education-line database 27 November 1998. http://www.leeds.ac.uk/educol/

Murphy, R.J.L., (1982) A further report of investigations into the reliability of marking of GCE examinations. The British Journal of Educational Psychology, 52, 58-63.

Polanyi, M., (1958), Personal Knowledge, London: Routledge.

Polanyi,M., (1962), Personal Knowledge, London: Routledge.

Senge, P.M. (1990) The fifth discipline: The art & practice of the learning organization (1993). London: Random House UK Ltd.

Shor, M., and Robson, R., (2000) A Student-Centered Feedback Control Model of the Educational Process. 30th ASEE/IEEE Frontiers in Education Conference

Torrance, H., (1994) Curriculum Assessment and Evaluation - Changing Conceptions and Practice, Paper prepared for the Association for the Study of Educational Evaluation in Southern Africa Conference, 6-8 July, 1994, Pretoria. http://www.aseesa-edu.co.za/currasse.htm

Wiliam, D. (1997) Construct-referenced assessment of authentic tasks: alternatives to norms and criteria. http://www.kcl.ac.uk/depsta/education/hpages/EARLI97.pdf

Wiliam, D. (1998) Construct-referenced assessment of authentic tasks: alternatives to norms and criteria. Paper presented at the 24th Annual conference of the International Association for Educational Assessment – Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, May 1998.

Wiliam D,. (2000). The meanings and consequences of educational assessments. Critical Quarterly, 42(1), pp105-127 (2000) http://www.aaia.org.uk/pdf/2001DYLANPAPER2.PDF