Connoisseurship, assessments of performance and questions of reliability

Paper presented at the 32nd Annual IAEA Conference, Singapore 2006

Jonathan H Robbins
The Talent Centre Ltd

Abstract
Increasing emphasis is being placed on the importance of personal qualities and attributes such as creativity, inter-personal skills, leadership, communication, or aesthetic awareness. These qualities and attributes are not amenable to current assessment processes that rely on specifications and the production of evidence or on more conventional types of examination. This paper describes applications of connoisseurship and construct referencing in assessing achievements of these or similar qualities through observed performance. Examples are drawn from assessments of performance in music and dance as well as from programmes concerned with emotional and behavioural development, employability and training for teaching and learning support. Questions relating to applications of connoisseurship and the reliability of assessment practices and results are considered. A brief description of methods used to express measures of reliability in aviation training and in precision engineering is given and their application to the reliability of assessments of performance considered. Results obtained over a three year period in the use of a method developed for monitoring assessor performance and the standardisation of results from assessments of performance by an awarding body in the United Kingdom is reported on together with some conclusions about its use and applicability in different settings.

Introduction

Increasing emphasis is being placed on the importance of personal qualities and attributes such as creativity, inter-personal skills, leadership, communication, or aesthetic awareness. These two illustrations: a Google search on the terms – 'recognise, creativity, leadership, communication' that generated over nine and a half million 'hits'; and the following extract from a speech by Sir Nigel Crisp entitled 'Nurse Leadership for the Future' to the Chief Nursing Officers Conference[1] on the 10 November 2005 may serve as simple examples:

> *During this phase of the reforms even more will be required of nurse leaders. Operating in an increasingly complex and competitive environment, nurse leaders must demonstrate higher level leadership, communications and management skills than before.*

These and similar qualities and attributes are not amenable to current assessment processes that rely on specifications, criterion referencing and the production of evidence or on more conventional types of examination. This is because as Gipps and Stobart (1996)[2] remind us '*as the requirements become more abstract and demanding, so the task of defining the performance clearly becomes more complex and unreliable*'. Difficulties with criterion referencing are widely recognised and the monograph by Glass(1997)[3] provides a valuable overview and discussion of these. One response has been the promotion of construct referenced assessment, Wiliam (1998)[4] as a more appropriate basis for judgement. This may be an improvement on criterion referencing described by Wiliam(2000)[5] 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.

Connoisseurship and Assessment

Connoisseurship involves expert norm-referenced judgements being made by a person who is recognised as having the knowledge and experience necessary to do so. Examiner, observer, rater and assessor are all terms that may be used in different settings to describe this person. In the United Kingdom the use of a connoisseurship model of assessment is widespread and has a long tradition of use, especially in the arts and for teacher assessment of coursework, portfolios and investigative or project work. 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). The term 'educational connoisseurship' is used by Eisner (1998a)[6] 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)[7] and (1998b)[8].

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. 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.

Connoisseurship Characteristic:	**Met 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.	Professional qualifications Music qualifications Examiner approval and registration Significant teaching experience Specialist knowledge of instrument examined Significant experience of the practice of music (e.g. as a player, composer, conductor)
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.	Music education within the tradition of graded examinations Further professional training as a musician Significant experience as a practicing musician and teacher over a long period Training as an examiner with 'shadowed' or parallel marking, mentoring or similar induction procedures
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.	Standards determined by published repertoire, descriptions of performance requirements, grade and boundary descriptions. No rank ordering of candidates No post-examination grading against results
Purposes for the assessment that are shared and agreed both within and outside of a community of practice.	Published syllabus, repertoire and marking scheme Entry for examination when deemed ready by teacher Mastery of candidate assessed against what has been previously published, prepared and practiced
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.	Systematic monitoring of examination results and examiner performance with feedback to examiner Appointment as an examiner requires proof of ability to mark consistently and comparably within stated bounds
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.	Award based only on the examiners judgement of the extent to which a stated threshold of performance is met on specific aspects of the examination and on the performance of the candidate as a whole.

Figure 1. Applications of connoisseurship assessments to Graded Examinations of Musical performance

Connoisseurship and Reliability

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. Understanding the relationship between a connoisseurship model of assessment and other forms of assessment is important to an understanding of what reliability means and to the development of ways to quantify reliability that are 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 can be visualised in the format shown in Figure 2. 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.

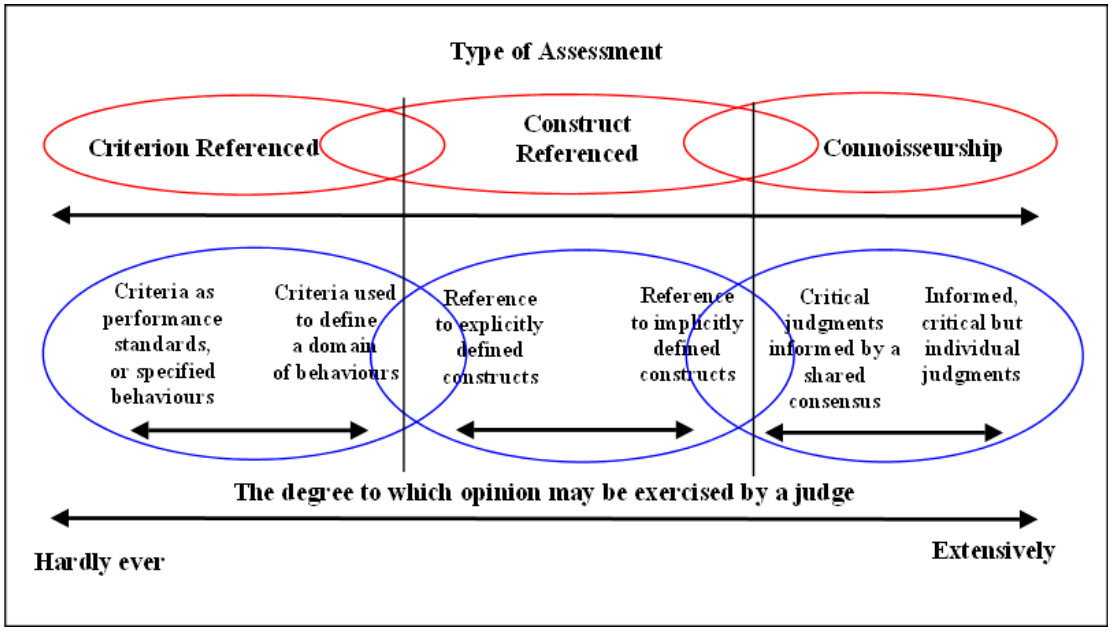

Figure 2. A continuum - Criterion referencing to connoisseurship.

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. Broadfoot (1998)[9] 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)[10] makes similar observations in a more general context. For assessments of performance, it is not whether one way of quantifying reliability is better than another, but rather a matter of selecting a way of stating to what extent a particular form of assessment is appropriate to its purpose and can be relied upon. To paraphrase the comment of Wiliam (1997 July)[11] about raters, 'to put it crudely it doesn't matter how reliability is calculated, only that what it tells us is relevant'.

Johnson and Goldsmith (1998)[12] and Holt, Johnson and Goldsmith (1998)[13] 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. Johnson and Goldsmith (1998)[14] suggest a 'multi-pronged' approach where the meaning of reliability as a measure of consistency, is extended to include properties of sensitivity and accuracy. Two methods are presented for assessing the reliability of observations, rater-referent reliability (RRR) and inter-rater reliability (IRR). 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. Statistical process control is widely used in precision engineering and is principally concerned with the stability of a process and the removal of variation. 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 (1995)[15] 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 connoisseurship models of assessment, reliability or the extent of 'sameness' may be described as the amount by which assessment decisions may 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'. A possibility arising from this is the application of the concept of a region of acceptably stable results. This requires the setting of upper and lower control limits or 'bounds' which control the extent to which results 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.

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 GCE 'A' Level or GCSE, 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 Cronbach's 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.

The concept of reliability as the stability of rater judgements, 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.

The stability of rater judgements is being investigated as part of ongoing research into the use of the Affirmative Assessment System™ for Graded Examinations of Dance (Classical Ballet, Tap, and Modern Jazz). The results of this research will be reported at a later date.

Address for correspondence: robbinsj@talent-centre.com
The Talent Centre Ltd Nottington Weymouth Dorset DT3 4BH England.
www.talent-centre.com

References

[1] Speech by Sir Nigel Crisp, Chief Executive, to the CNO Conference, 10 November 2005 Published: 10 November 2005. Accessed 01/03/2006 http://www.dh.gov.uk/NewsHome/Speeches/SpeechesList/SpeechesArticle/fs/en?CONTENT_ID=4126396&chk=bp4n3/

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

[3] 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

[4] 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.

[5] Wiliam, D., (2000) The meanings and consequences of educational assessments *Critical Quarterly*, **42**(1), pp105-127 (2000)

[6] 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

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

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

[9] 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

[10] 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

[11] Wiliam, D., (1997, July) *How to do things with assessments: illocutionary speech acts and communities of practice.* Paper presented at 21st Conference of the International Group for the Psychology of Mathematics Education Discussion Group on Assessing Open-Ended Work in Mathematics held at Lahti, Finland. London, UK: King's College London School of Education.

[12] 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

[13] 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

[14] ibid

[15] Gipps, C.V. (1994) *Beyond Testing: Toward a theory of educational assessment (1995).* London: The Falmer Press.