# Holistic judgement of a borderline vocationally-related portfolio: a study of some influencing factors

**Martin Johnson** Research Division

## Background

Literature suggests a number of background issues that might be pertinent to this area of work. The assessment of a large portfolio of mainly textual evidence demands an assessor to accommodate a great deal of information. It has been suggested that assessors' initial comprehension of a text is an important consideration (Huot, 1990b; Sanderson, 2001). This comprehension process is influenced by the linear nature of the reading process which leads to the gradual construction of a mental representation of the text in the head of the reader (Johnson-Laird, 1983). Another cognitive factor to consider relates to the use of 'generic' phrases in assessment criteria. Oates (2004) argues that these can exact a large cognitive demand on assessors if their use is dispersed across different contexts and/or assessors do not encounter the descriptors very frequently. Finally, it is important to consider the value system within which the reader/assessor is located and which might affect their thinking. Sanderson (2001) suggests that the social context of the assessor is important to consider since it recognises their participation in a community of practice (Wenger, 1998) and constitutes an 'outer frame' for their activity.

It is also important to consider how assessors integrate and combine different aspects of an holistic performance into a final judgement. Most study findings appear to support the suggestion that between-marker reliability is greater for analytic scoring methods, where individual scores are given across multiple dimensions, rather than holistic scoring methods, where a single score is given across multiple dimensions (Breland, 1983; Huot, 1990a; Johnson *et al*., 2001). Laming (2004) argues that this is because linear combinations of individual diagnostic signs have greater accuracy than more strictly holistic judgements because they use an arithmetic basis. Other studies also discuss this problematic issue, suggesting that overall judgement is often based on the cumulative weighting and combination of cues found within a performance and that these weightings might vary (Vaughan, 1991; Einhorn, 2000; Elander and Hardman, 2002).

The recent works of Engeström (2001) and Wenger (1998, 2000) have been very influential in terms of recognising the importance of socio-cultural influences for understanding individual behaviours. This has implications for inter-assessor consistency because it suggests that there is a need to reflect on the role that the social dimension plays in assessment judgements including the potential existence of differing interpretations and standards between assessors.

Investigating socio-cultural influence on assessor consistency has implications for the research method chosen. Whilst socio-cultural theory suggests that human behaviour needs to be understood in the context of the interactions between the characteristics of people and their environments, Rapport *et al*. (2004) characterise 'scientific' knowledge as being independent of time and place with observed variations explained through relevant theory. Popular cognitive research methods, such as Kelly's Repertory Grid (Kelly, 1955) or Verbal Protocol elicitation techniques often conform to this experimental scientific model, focussing on individualised data collection whilst potentially overlooking the influence of the social environment on those elicitation processes. On the other hand, descriptive qualitative methodologies, such as observation and interview techniques, can consider the interaction of both social and individual elements. Bronfenbrenner (1979) argues that understanding might be progressed by uniting the schismatic experimental and descriptive psychological traditions through designing research studies which combine ethnographic and more 'controlled' methods.

This present study attempted to accommodate both of these perspectives by using an integrated approach to data collection. It sought to explore issues of consistent assessor judgement by gathering data about individual assessors' cognitive activity as well as the socio-contextual features in which their practices were undertaken.

## Method

This study was set in the context of the OCR *Nationals* in Health and Social Care (Level 2). This qualification was chosen because assessors use an holistic, best fit grading model, organised into a number of Assessment Objectives (AO) to judge portfolios of students' work. Six assessors were involved; four assessors (M1-M4) were Visiting Moderators for the qualification and the others (T5 and T6) were experienced OCR *Nationals* course tutors.

In order to investigate the factors that they attended to during the assessment process the assessors were asked to 'think aloud' whilst they judged a Unit 10 (*preparing to work with people with disabilities*) portfolio which had already been identified as having pass/merit borderline characteristics. This commentary, taken to be a partial record of the features that the assessor attended to during the assessment activity, was transcribed into a verbal protocol and analysed with qualitative text analysis software.

A modified Kelly's Repertory Grid (KRG) interview technique was also used to gather data about different assessors' perceptions of constructs within the same assessment criteria. This activity focussed on the grading criteria for Unit 1 (*preparing to give quality care*). The theory underpinning this method is based on Kelly's model of Personal Construct Psychology (Kelly, 1955), which suggests that individuals possess a constructed version of their world based on their experience. This construction comprises personally held bi-polar mental constructs which can be elicited through KRG techniques. This method asks individuals to verbalise salient differences and similarities between triads of objects or 'elements'. These salient features and patterns anchor ends of bi-polar constructs along which individuals can place other different objects or 'elements'. This method was used to elicit the constructs that assessors

perceived within the grading criteria for each Unit 1 AO. These constructs were then related to their judgements during the portfolio assessment exercise in order to explore whether data about construct elicitation and grading criteria interpretation could shed light on issues of consistent judgement-making.

Qualitative contextual data were collected through observations of three moderation visits to schools and colleges in different parts of England involving three of the assessors in this study. These visits enabled case study evidence to be collected through structured field notes to record details about the different sections of the moderation meetings, the amount and diversity of work covered, and contextual working information. These data also fed into the drafting of questions for the next level of data collection where each assessor was interviewed following the portfolio assessment activity. These semi-structured interviews gathered information about assessors' professional background in order to highlight any potential influences upon their assessment practices.

The final stage of analysis involved the integration of evidence from the different sources of data collection. In the first instance this entailed isolating the salient features identified within the VP and KRG data and cross-referencing them to the features identified in the observation and interview data to identify any linkages and patterns. It needs to be acknowledged that this process contained a subjective quality. It ignored some of the individual micro level linkages that might have been discernible through a more fine grained analysis in order to focus on triangulation at the macro level to identify the larger themes within the data.

**Table 1: Frequency of assessor judgements at each grade**

| | | *Fail* | *Pass/Fail* | *Pass* | *Merit/Pass* | *Merit* |
|---|---|---|---|---|---|---|
| AOs | 1 | | | 1 | 1 | **4** |
| | 2 | 1 | | **3** | | 2 |
| | 3 | 2 | | **4** | | |
| | 4 | 1 | | 3 | | **2** |
| | 5 | 1 | 1 | **2** | 1 | |
| | 6 | 1 | | **2** | 2 | |

*Bold indicates agreement with original portfolio assessment

**Table 2: Mean assessor agreement levels**

| | *M1* | *M2* | *M3* | *M4* | *T5* | *T6* |
|---|---|---|---|---|---|---|
| M1 | — | 0.17 | 0.8 | 0.5 | 0.67 | 0.25 |
| M2 | 0.17 | — | 0.33 | 0.67 | 0.67 | 0 |
| M3 | 0.8 | 0.33 | — | 0.67 | 0.8 | 0 |
| M4 | 0.5 | 0.67 | 0.67 | — | 0.8 | 0 |
| T5 | 0.67 | 0.67 | 0.8 | 0.8 | — | 0.25 |
| T6 | 0.25 | 0 | 0 | 0 | 0.25 | — |

## Findings

Although this study was not solely concerned with gathering reliability data, differences between the frequencies between assessors' judgements at different grades during the assessment exercise suggested that there was potential for further investigation of the factors that might have affected their judgements (Table 1).

T5 exhibited the greatest overall degree of agreement with other assessors (Table 2). T6 was the most severe assessor. M3 and M2 had the highest and lowest respective levels of agreement with the most senior assessor (M1).

It is important to acknowledge two potential factors that might have influenced the assessors' judgements: it is possible that the think aloud data collection method might have influenced the assessment process; and two of the assessors (M2 and T5) suggested that they lacked familiarity with the particular unit being assessed since both lacked teaching experience of this particular unit, although they both moderated the unit.

The areas of high shared focus in this study were found around areas of the portfolio that were 'signposted' by textual devices such as clear headings and titles. This search for evidence was itself clearly structured by the grading and KUS (knowledge, understanding and skills) assessment guidance as assessors tended to navigate the portfolio by searching for performance evidence in a similar order. Those assessors who rated Unit 10 AO2 most severely were more likely to attend to features embedded within the text and away from the common areas of attention around the 'signposts', and particularly further on in the portfolio.

There were very clear areas where assessors' comments suggested that they were attending to similar ideas and basing their decisions on similar frameworks. In some Unit 10 AOs it was apparent that fundamental values influenced assessors' practice. In AOs 5 and 6 the dominant influence of 'care values' was evident whilst in AO3 it was 'application'. A 'positive assessment' culture also appeared to pervade the practices of these assessors where they looked to highlight the achievement of the learner. This contrasts with some of the practices identified in other areas of general/academic assessment (Sanderson, 2001; Crisp and Johnson, 2007). These positive assessment practices appear to be underpinned by a strong desire to motivate learners, which was a theme clearly articulated by different assessors during interview. One potential concern that this raises is that assessors might tend to give learners the benefit of any doubt when they are in two minds about the quality of a performance, particularly if schools/colleges fail to prepare their students with appropriate tasks or guidance. KRG analysis also alluded to the presence of shared values through the identification of four 'core' constructs across the different Unit 1 AOs. These constructs were: *application* (4 AOs); *description or account quality* (4 AOs); *sources* (4 AOs); and *example use* (3 AOs). Of these, *application* was notable because assessors consistently weighted it very highly, suggesting it to be a very strong core feature of assessment for these judges.

There was also evidence that assessors' values might have affected their practice in other common ways. Verbal protocol analysis showed that some elements within the grading criteria tended to be attended to more than others, perhaps reflecting the value placed on them by the assessors. Assessors appeared to inherently respect having another competent professional to judge the student's proficiency within a contextualised learning environment. In this study assessors alluded to some of the potential problems that this might lead to, particularly when assessors are not given the right degree of information or where it isn't provided in a useful format. The verbal protocol data also suggested evidence of an assessor using the student performance on the practical

task AO to justify her final judgement for the whole portfolio.

There was also evidence of discrepant practice between assessors. Verbal protocol evidence showed that some assessors adopted a linear strategy to combine several equally weighted factors within AOs, whilst others assigned some performance factors unequal weighting. One example of this was Unit 10 AO4 which contained a third party witness statement suggesting that the student's performance warranted a pass grade. Two assessors appeared to assess this practical task evidence in a linear fashion, balancing it equally alongside other AO evidence, and reaching a 'merit' grade overall. For the other assessors it appears that the witness statement might have been a major influencing factor on their final evaluation which suggested a 'pass' grade overall.

Assessors elicited 131 KRG constructs over the six AOs. The most senior assessor (M1) elicited more constructs on average per AO (7.8) than either the other moderators (4.9) or the tutors (5.0), and t-test analysis showed that this difference was significant ($t = 8.16$, $p < 0.01$). Despite this level of verbalisation the most senior assessor found it difficult to separate these constructs into component aspects across the borderlines, potentially signifying the highly tacit nature of important features of this knowledge.

KRG analysis also identified some potentially problematic issues around lexical interpretation. Some of these clustered around 'construct fusion'. It was possible to find instances where assessors felt that the concepts of 'quality' and 'quantity' had become fused as they progressed through the grade descriptors, such as where descriptors used adjectives relating to the quality of a concept (e.g. *simple* or *basic*) alongside adjectives relating to their quantity or existence (e.g. *some*) (Unit 1: AO1 and AO3). Some assessors also perceived that some qualitative aspects of the descriptors lacked discrimination or appeared to overlap. Assessors sometimes expressed difficulty in separating some of the descriptive qualities within the criteria because the terminology failed to adequately describe differences as they understood them. For example, '*organising information appropriately*' (Unit 1: AO2 pass) might also involve it being '*clear, accurate and detailed*' (Unit 1: AO2 merit), or, assessors might expect a '*basic*' understanding of an issue to be also '*sound*' (Unit 1: AO2 and AO3).This issue also linked to the parallel finding in the interview data where some assessors suggested that they knew where to locate commonly agreed meanings for important words, although the location of this resource varied. This aspect of consistent application, and the potential for misaligned understandings, also resonates with other anecdotal data from the early set up stages of the project which suggested that tutors in schools/colleges sometimes assign their own common 'in house' meanings to descriptor terminology.

The verbal protocol data appear to suggest that assessors might find it difficult to focus on particular performance elements in isolation when reading through work. This highlights a potentially central tension for these vocationally-related assessors who have a strong philosophical attachment to holistic assessment. It is also possible to suggest that holistic assessment might allow assessors to avoid areas of an assessment scheme where there is a lack of clear understanding about the meaning of certain criteria. Although this can lead to better levels of consistency it potentially masks a problem nested within the assessment criteria and which needs to be dealt with.

The observation and interview data identified some key pressures relating to the workloads of moderators. For example, they were under pressure to complete the moderation paperwork during their school/college visit whilst at the same time fostering and maintaining positive links with their hosts to support their ongoing development. These demands are potentially contradictory, with the external validity of the qualification at risk if the balance is not correctly struck.

Another interesting issue found in the interview data was the existence of networks beyond the bounds of this qualification that might have had an effect on assessor practice. Assessors 1 and 3 exhibited the highest levels of inter-assessor agreement in the portfolio assessment exercise and they also shared some common frameworks which did not necessarily overlap with other assessors. These shared frameworks included an understanding that 'evaluation' required 'justification', 'synthesis' acted as a key quality indicator, and the use of a linear rather than a holistic method when accumulating different elements into a final judgement. It might be tentatively suggested that these similarities might have been reinforced by the close connection that these assessors had through their contact through moderation work in another Health and Social Care qualification. Acknowledging the possibility that this external link might overlap into the *Nationals* environment is important because it represents one of the networks (and related tools) that might exist and to which some assessors have restricted access.

## Implications

The manner in which the assessors balanced some of the information when reaching a judgement appeared to interact with their underlying values. It could be important for these values, of which 'application' and 'generality and synthesis' appear to be core elements, to be elicited and acknowledged. This might help to undermine the often tacit nature of vocational values and help to promote a common codified framework as a basis on which to discuss interpretations of performance evidence.

There was evidence that some assessors tended to combine performance features in a linear fashion whilst others allowed certain features to dominate their overall judgements. Theory suggests (e.g. Laming, 2004) that the linear method should promote better consistency levels but it is important to explore why some assessors might value particular aspects of performances more than others. Discussion about the appropriate way to balance such features could form an important part of the initial training for assessors new to the qualification and their subsequent moderation visits.

The KRG data suggest that the more experienced assessors (who elicit the greatest number of constructs) might find it most difficult to break down their judgement-making processes. This might represent a challenge to the induction of new assessors.

Concerns about 'construct fusion' require a careful evaluation of the grading criteria to trace the development of constructs through boundaries in order to identify where aspects of concept quality or quantity might overlap. The KRG methodology might be a useful technique for such an activity. A consequence of this process would also be to allow training and moderation visits to draw assessors' attention to this potential problem within the criteria so that they can be aware of it when making judgements. This feature could also factor into any future assessment criteria development programmes.

Consistent lexical interpretation could be further supported by having a clearly referenced resource available for qualification users that defines the meaning of key terminology (e.g. the terms '*range*' or '*simple*' and '*detailed*'). This would reinforce the messages given at training sessions where literal explanations of terminology might be given to new qualification users. This could also be followed up through discussions

around the meanings of key terms during moderation visits.

Assessors sometimes expressed difficulty in separating some of the descriptive qualities within the criteria because, from their perspective, the terminology failed to adequately illustrate differences between the qualities of different performances. This implies that the language used either did not conform to discrete categories or had some overlapping qualities (e.g. '*clear*'/ '*accurate*'/ '*appropriate*'/ '*detailed*' or '*basic*'/ '*sound*'/ '*high*'), that made it difficult for assessors to fit some performance characteristics to the criteria. Although caution needs to be expressed about making assessment criteria more lengthy (Wiliam, 1998; Wolf, 1995), resolving this issue might involve clarifying the values implicit in the descriptor terminology, perhaps through exemplification, and connecting these meanings through effective communication procedures with assessors' expectations about performance quality. This implies a need to engage assessors in discussions about those aspects of language that they feel hinder their ability to discriminate between performances and to use this as an opportunity to arrive at agreed meanings.

**References**


Breland, H. (1983). *The direct assessment of writing skill: a measurement review*. Technical Report No. 83–6. Princeton, NJ: College Entrance Examination Board.

Bronfenbrenner, U. (1979). *The ecology of human development: experiments by nature and design*. Cambridge, MA: Harvard University Press.

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–962.

Einhorn, H. J. (2000). Expert judgement: some necessary conditions and an example. In: T. Connelly, H. R. Arkes & K. R. Hammond (Eds.), *Judgement and decision making: an interdisciplinary reader*. 2nd edition, 324–335. Cambridge: Cambridge University Press.

Elander, J. & Hardman, D. (2002). An application of judgement analysis to examination marking in psychology. *British Journal of Psychology*, **93**, 303–328.

Engeström, Y. (2001). Expansive learning at work: toward an activity theoretical reconceptualization. *Journal of Education and Work*, **14**, 1, 133–156.

Huot, B. (1990a). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, **60**, 237–263.

Huot, B. (1990b). Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication*, **41**, 210–213.

Johnson, R. L., Penny, J. & Gordon, B. (2001). Score resolution and interrater reliability of holistic scores in rating essays. *Written Communication*, **18**, 2, 229–249.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

Kelly, G. A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Laming, D. (2004). *Human judgment: the eye of the beholder*. London: Thomson Learning.

Oates, T. (2004). The role of outcomes-based national qualifications in the development of an effective vocational education and training system: the case in England and Wales. *Policy Futures in Education*, **2**, 1, 53–71.

Rapport, F., Wainwright, P. & Elwyn, G. (2004). "Of the edgelands": broadening the scope of qualitative methodology. *Journal of Medical Ethics; Medical Humanities*, **31**, 37–42.

Sanderson, P. (2001). *Language and differentiation in Examining at A Level*. PhD thesis, University of Leeds.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, **7**, 2, 225–246.

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, Barbados.

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




# Annotating to comprehend: a marginalised activity?

**Martin Johnson** Research Division **and Stuart Shaw** CIE Research

## Introduction

One of the important premises underlying this article is that the cognitive processes involved in reading can play a significant role in assessment judgements. Although we acknowledge that not all assessments of performance rely on assessors appraising written texts, many tests use written evidence as an indicator of performance. As a result, it is important to consider the role of assessors' comprehension building when reading candidates' textual responses, particularly where candidates are offered a greater freedom in determining the form and scope of their responses.

Crisp and Johnson (2007) note that it is common practice for examiners to annotate scripts when marking. This convention is formalised in the Qualifications and Curriculum Authority (QCA) code of practice (QCA, 2007) which stipulates that a second assessor needs to see any annotations made by a first assessor to gain a full and clear understanding of whether the marking criteria have been applied as intended. Beyond this formalised role, annotation might perform a more general and less formalised function in individual reading comprehension building processes.

Sources (Weiner and Simpson, 2005; Merriam-Webster, 2005) suggest that the definition of the word 'annotation' is to be found in the