Assessing at the borderline: Judging a vocationally related portfolio holistically

Martin Johnson
University of Cambridge Local Examinations Syndicate

This study investigated the cognitive strategies that underpin assessors' holistic judgments of a school-based vocationally-related portfolio performance. Using a portfolio already identified as containing borderline qualities, quantitative data were gathered about features that six assessors attended to as they holistically evaluated the portfolio. This information was gathered through verbal protocols and supplemented with information from a modified Kelly's Repertory Grid interview technique (Kelly, 1955). This elicited assessors' perceptions about the characteristics of the assessment criteria, allowing the influence of each factor to be ranked.

Another objective was to collect qualitative data about the socio-contextual features in which the assessors' practices were situated. The study uses Activity Theory to explore the relationship between the factors that potentially influence assessors' judgments, suggesting a theoretical position that assessor judgments are influenced or framed within the context of their experience and differing perspectives.

Introduction

Consistency of assessor judgment is a key concern for those charged with accrediting learning outcomes. This is partly because these judgments can have very real consequences for learners' future employment or further education chances. Concerns about the relationship between vocationally-related assessment and context are well recognised. Literature suggests that the consistency of assessor judgments can be affected by the breadth of learning contexts (Johnson, 2007; Curtis, 2004; Hager, 2004) and concomitant learner and assessor experiences.

The provision of assessor support, through common training and well designed assessment criteria can help to mitigate some of these concerns. It is also important to understand how experts make judgments about learners' performances since this might help to make the assessment system more transparent and able to justify claims of fairness. This is particularly the case for holistic assessments where assessors might attend to a variety of factors in different contexts when forming a judgment.

This study focuses specifically on how six assessors holistically judge a portfolio of evidence containing borderline pass and merit characteristics. It uses an integrated approach to collect data about the assessors' cognitive activity as well as the socio-contextual features in which their practices are undertaken. The assessors all typically worked with the chosen qualification in different contexts; four assessors were visiting moderators and two assessors were course tutors.

The study involved the assessors using a 'think aloud' method whilst judging a Health and Social Care portfolio containing material from Unit 10 *(preparing to work with people with disabilities)*. The second cognitive approach adopted was to use a modified Kelly's Repertory Grid (KRG) interview technique to gather data about different assessors' perceptions of constructs within the assessment criteria for Unit 1 *(preparing to give quality care*). These cognitive approaches were augmented by qualitative data collected from observations of three separate moderation visits to schools and colleges in different parts of England. These observations also fed into the drafting of questions for the next level of data collection where each assessor was interviewed following the portfolio re-assessment activity. These semi-structured interviews gathered information about assessors' professional background details in order to highlight any potential influences upon their assessment practices.

Although it needs to be the recognised that the methods used for gathering socio-contextual data in this study are partial, and would only be expected to offer a limited insight into any differing perspectives, it is quite noticeable that the assessors in this group shared many key values.

Holism, atomism and assessing performance

Qualifications in the UK can be broadly categorised into three types: vocational, general vocational/vocationally-related or general/academic. Table 1 describes some of the differences between the qualification types. Although this characterisation is an oversimplification of some of the finer details of the current system, it is a useful basis for discussion. Each of the different qualification types typically uses different forms of evidence with different structures in place to support assessor consistency. One noticeable difference is between their 'assessment density', a concept that involves the frequency with which assessors see the same sorts of performance evidence in similar contexts. This concept might help to explain why there have been relatively few investigations into the reliability of vocational assessment models in the UK, with notable exceptions being Murphy et al. (1995) and Eraut, Steadman, Trill and Parkes (1996).

This study is set in the context of a vocationally-related *Nationals* Health and Social Care qualification developed by the Oxford, Cambridge and RSA (OCR) examination board. OCR is one of the three largest providers of general and vocational qualifications in the UK. According to the OCR centre handbook (OCR, 2006) the *Nationals* qualification has been developed to recognise candidates' skills, knowledge and understanding of the health and social care sector and the settings, job roles, principles and values involved. The qualification is usually delivered to 14-19 year olds in schools and colleges, with additional opportunities for them to demonstrate their learning in applied situations.

In order to achieve certification the candidates must achieve a minimum pass grade for four mandatory units and two optional units. A portfolio of performance evidence for each unit is graded holistically, as pass, merit or distinction, recognising that candidates may perform better in meeting the requirements of some objectives more than others. All

units are assessed by tutors in their place of learning and externally moderated by an OCR Visiting Moderator.

Table 1: UK qualification types

Qualification type	Assessment activity	Structures to support consistent assessment	Example
Vocational	Focus on performance outcomes: may include observation of the candidate, examination of a product, witness testimony etc.	Internal and external verification procedures	National Vocational Qualifications (NVQ)
'General vocational'/ vocationally-related	Assessments of portfolio evidence in different units: judges evaluate performances against the unit assessment objectives in a pass, merit or distinction fashion.	Tutor CPD; Assessor training; Internal/external moderation.	OCR *Nationals*
General/ academic	Marking of terminal written examinations	Coordination and standardisation.	General Certificate of Secondary Education (GCSE)

The assessors in the study were all experienced practitioners with current teaching experience in the Health and Social Care field. One assessor was the national Chief Coordinator of the qualification who also taught part-time, whilst the other assessors held management roles concerning the Health and Social Care provision within their own teaching institution. The assessors had worked at their current institution between 3 and 36 years (mean 15.8 years) and had held their current position between one and 18 years (mean 9.2 years). Alongside their teaching commitments the assessors also had strong links to applied care related activities. Five of the assessors were involved currently or in the past with aspects of nursing, voluntary homeless work, counselling and family social work units.

Although details about the student who authored the portfolio are not available it is worth reflecting on some circumstantial information that might be relevant. In line with other recent developments in UK vocational policy the OCR *Nationals* might be seen to carry an element of social purpose; aimed at individuals who are not succeeding at school (Stasz & Wright, 2004). Evidence for this suggestion is reflected in some of the assessors' comments on the inherent potency of applied learning whilst also talking at length about how the qualification assessment model suited the learners; suggesting a number of reasons why its assessment model was compatible with their shared objective of motivating the learners with whom they were involved. These features included its lack of testing, its holistic approach to judging performance, and a positive assessment approach to recognising students' achievements.

The assessment of a portfolio of mainly textual evidence demands an assessor to accommodate a great deal of information. Research literature suggests that assessors'

initial comprehension of this text is an important consideration (Sanderson, 2001; Huot, 1990). This in itself might be problematic since some theorists argue that the linear nature of the reading process leads to the gradual construction of a mental representation of the text in the head of the reader (Johnson-Laird, 1983). Although this mental model might in itself be difficult to appraise objectively, it might be partially evident through the outwardly visible behaviour of the reader. Through considering the features of the text to which they attend it might be possible to infer the textual features that they are considering during their process of meaning making.

Another important consideration is the value system within which the reader/assessor exists and which affects their thinking. Sanderson (2001) suggests that the social context of the assessor is important since it provides an 'outer frame' for their practice and involves their participation in a community of practice (Wenger, 1998). The assessors in this project frequently worked with holistic performance criteria to evaluate performances, and to some extent this feature might be seen to represent something that they value. If this is the case there is a potential tension between the desire to make holistic judgments based on reading/comprehension structures that are linear and in some senses atomistic in nature.

Finally, it is important to consider how assessors integrate and combine different aspects of an holistic performance into a final judgment. Laming (2004) cites a number of judgment studies highlighting the difficulties that clinicians experience when combining different observations to reach a conclusion. He argues that linear combinations of individual diagnostic signs have greater accuracy than more strictly holistic judgments and that 'the reason [for this] is that the combination of diagnostic signs by the clinician is no better than qualitative; it parallels the ordinal quality of human judgment. But the linear combination uses arithmetic, and is therefore superior' (2004, p.64). Other studies also highlight this problematic area, suggesting that overall judgment is often based on the cumulative weighting and combination of cues found within a performance and that these weightings might vary (Elander & Hardman, 2002; Einhorn, 2000; Vaughan, 1991).

Assessor consistency: A sociocultural perspective

The recent works of Engeström (2001) and Wenger (1998, 2000) have been very influential in terms of recognising the importance of sociocultural influences for understanding individual behaviours. Considerations of inter-assessor consistency also need to reflect on the role that the social dimension plays in assessment judgments since this might help to explain the existence of differing interpretations and standards between assessors.

Wenger's notion of 'Community of Practice' has been used to describe groups of people working together in assessment communities (Baird, Greatorex & Bell, 2004; Price, 2005; Crisp & Johnson, 2007). Assessment literature suggests a number of areas where the crucial factor of shared communication might falter leaving space for misaligned standards, differing interpretations and diminished between-assessor consistency. One of these areas relates to the 'plasticity' of assessment criteria. Observers suggest that

assessment criteria are complex objects which cannot convey every possible meaning and can inevitably leave room for individual assessor interpretation (Saunders & Davis, 1998; Wolf, 1995; Wiliam, 1998). Saunders and Davis (1998) also list a number of other potentially problematic issues which might differentially influence individual assessors. These include; varying affective reactions to work presentation, assessor fatigue, and speed or lack of time. Literature also adds differential assessor experience levels (Johnson, Penny, Gordon, Shumate & Fisher, 2005) and order effects (Spear, 1997) to this collection.

This concern has important consequences for the activity of applying consistent assessment standards since Laming (2004) argues that everybody makes judgments in the light of their past experience and in instances where judgments are uncertain "past experience enters like air rushing into a vacuum" (2004, p.164). Dunn, Parry and Morgan (2002) highlight the importance of recognising the role of values across a community of individuals. They suggest that assessors' consistent interpretation and application of criteria are underpinned by common norm values. Some theorists suggest that understanding norms requires a sociocultural perspective since interpretations are contextually constituted and cannot be divorced from the value-bases which interpreters bring with them (Shay, 2005). In this conceptualisation differences between assessors are not 'error', but rather the inescapable outcome of the multiplicity of perspectives that assessors bring with them.

Finally, Beckett and Hager (2002) support the notion of 'embodied assessment'. They argue that practitioners in different situations have different amounts of time to reach decisions and that the expectation of discretionary judgments in the midst of fluctuating situations is a fundamental element that demarcates professions from each other. Consequently, assessment practice is more than a 'technique'; it involves a body of knowledge and a capacity to make judgments. Furthermore, professional judgments are holistic (integrating cognitive, attitudinal and emotional characteristics), and socially shaped (reflecting involvement in communities of practice and taking into account the specific characteristics of the situation in which they are made).

Mixing cognitive and sociocultural theories: Implications for method

'Scientific knowledge' can be characterised as being independent of time and place; with variations being explained through relevant theory (Rapport, Wainwright & Elwyn, 2005). Popular cognitive research methods, such as Kelly's Repertory Grid (KRG) or Verbal Protocol (VP) elicitation techniques often conform to this model, focusing on individualised data collection.

Nasir and Hand (2006) point out that sociocultural theories contrast with many cognitive psychological perspectives by locating the fundamental unit of analysis for the examination of human behaviour as activity, or cultural practices. One influential strand of theory is Activity Theory, which builds on Vygotsky's (1978) work recognising the influence of culture on individual actions. Engeström and others have extended Vygotsky's mediated action model to incorporate other important social and structural

features. This extended 'activity system' (Figure 1) explains change in terms of the evolving relationships amongst participants and the other elements of their environment. According to Kaptelinin et al. (1999) this system is a general conceptual approach rather than a highly predictive theory. It supports a dynamic model of analysis based around an object-oriented community which has multiple points of view, traditions and interests. It explains shifts in the understandings of its members through a dialectic process, focussing simultaneously on individuals and their community around an object (eg,. an assessment task) and the tools that they employ (eg, assessment criteria). Moreover, by foregrounding the multiple perspectives that exist around a singular activity it can help to identify sources of tension and conflict which drive change within the system. Such an analysis might help to explain why the application and interpretation of assessment criteria might vary at different locations within the community. Change or dissonant outcomes often occur where the system encounters internal contradictions, when new elements are introduced, or when alternative perspectives and practices are incorporated.

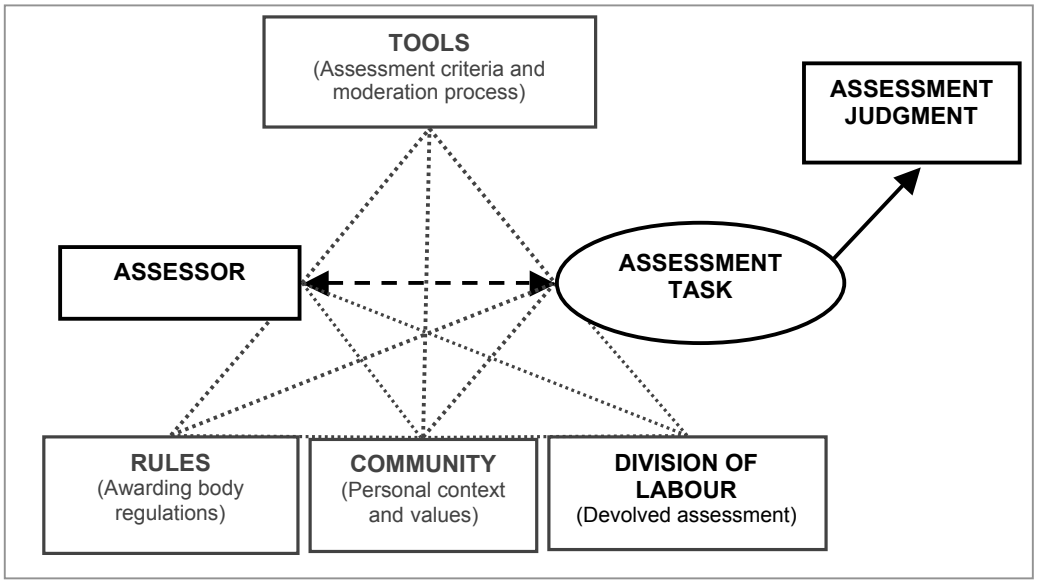

Figure 1: The structure of an assessment activity system (adapted from Engeström, 1987)

This theoretical perspective requires a qualitative methodology which can consider the interaction of both social and individual elements since methods aimed at eliciting cognitive data will potentially fail to capture the influence of the social environment on those elicitation processes. Recognising that the adoption of restricted research methods can limit the outcomes Bronfenbrenner (1979) argues for the use of mixed ethnographic and more 'controlled' research methods. He suggests that human actions need to be understood in the context of the interactions between the characteristics of people and their environments, therefore emphasising the importance of studying the environments

in which people behave and unifying the schismatic experimental and descriptive psychological traditions.

A mixed-method or integrated approach balances the strengths and weaknesses of the different methodologies. It enables triangulation by eliciting supplementary data (Pope & Mays, 1995) and allows the traditional scientific paradigm used for investigating phenomena in the physical world to be complemented by methods that are particularly suited to studying human beings in the social world (Watkins-Goffman, 2006).

Method

This present study uses an integrated approach to collect data about assessors' cognitive activity as well as the socio-contextual features in which their practices are undertaken. This forms the basis of a theoretical position suggesting that assessor judgments might be influenced or framed within the context of their experience which might help to explain any observed differences in perspective.

The project has two broad research questions.

- Which elements of holistic descriptors do assessors attend to when making judgments about borderline portfolio evidence?
- Which issues potentially affect the consistency of judgments made by different assessors who are working in diverse situations such as schools and colleges?

In order to answer these questions six assessors were asked to 'think aloud' whilst they judged an *OCR Nationals* Health and Social Care Level 2 portfolio. The assessors involved in the study all worked with the qualification in different contexts. Four assessors (M1-M4) were all visiting moderators for the qualification, with M1 being the most senior. The other assessors (T5 and T6) were *OCR Nationals* course tutors.

The portfolio chosen for this study contained material from Unit 10 *(preparing to work with people with disabilities)*, and was chosen because it had previously been considered by the most senior moderator to contain a variety of pass/merit borderline characteristics within its six units. The assessors used the usual qualification grading criteria, which is organised into six Assessment Objectives (AOs), and the assessors used the criteria in a holistic 'best fit' manner. This process allowed a verbal protocol (VP) of each judge's think aloud utterances to be gathered. This method has been used in a number of other research studies that have looked at assessor practices (eg, Milanovic *et al*., 1996; Suto & Greatorex, 2006).

The second cognitive approach adopted was to use a modified Kelly's Repertory Grid (KRG) interview technique to gather data about different assessors' perceptions of constructs within the assessment 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). This suggests that individuals possess a constructed version of their world based on their experience and that this comprises of personally held bi-polar mental

constructs. KRG techniques usually elicit these constructs by presenting an individual with triads of objects or 'elements' and asking them to identify any important ways in which two elements are viewed as similar to each other but different from the third. An individual's responses, based on the salient features and patterns that they perceive, anchor ends of a bi-polar construct along which the rating of different elements can be made. This method was used to elicit the constructs that assessors perceived to exist within the grading criteria.

The second research question involved researchers collecting qualitative data whilst observing three separate moderation visits to schools and colleges in different parts of England. These visits were conducted by two experienced visiting moderators and included a college where one of the tutors involved in this project was employed. These visits enabled case study evidence to be collected, including contextual and environmental features of the visits. This was achieved by non participant observation where researchers took 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. This data also fed into the drafting of questions for the next level of data collection where each assessor was interviewed following the portfolio re-assessment activity. These semi-structured interviews gathered information about assessors' professional background details 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, choosing to ignore some of the individual micro level linkages that might have been discernable through a more fine grained analysis, instead focussing on triangulation at the macro level to identify the larger themes within the data.

Findings

The different assessors' judgments during the re-assessment exercise showed some disagreement, including disagreement across the Pass/Fail boundary (Table 2). Two AOs (AO2 and AO4) were characterised by a spread of assessment judgments across two definite boundaries (fail/pass and pass/merit), whilst two (AO1 and AO3) had judgments spread over only one boundary (fail/pass or pass/merit).

It is important to interpret these data with some methodological caution since at least two factors might have influenced assessors' judgments. Firstly, it is possible that the think aloud data collection method used during the portfolio re-assessment exercise might have influenced the assessment process. Secondly, two of the assessors lacked teaching experience in the particular unit being assessed, although both had experience of moderating the unit.

Table 2: Assessor judgments

	M1	M2	M3	M4	T5	T6
AO1	**5**	**5**	**5**	**5**	4	3
AO2	**3**	5	**3**	5	**3**	1
AO3	1	**3**	**3**	**3**	**3**	1
AO4	**5**	3	**5**	3	3	1
AO5	**3**	1	**3**	4	2	-
AO6	**3**	1	**3**	4	4	-

1 = Fail; 2 = Pass/Fail; 3 = Pass; 4 = Merit/Pass; 5 = Merit

* Bold indicates agreement with original assessment

Holistic descriptors used in making judgments

VP data: AO specific features

On average, 215 codes were used when analysing each assessor's verbalisations. T-test analysis showed that Assessor 2 was assigned significantly fewer codes (91) and Assessor 4 was assigned significantly more codes (312) than other assessors across all AOs ($t = 6.32$, $p < 0.001$). This suggests that they were verbalising the things to which they were attending to a significantly greater or lesser extent than the other assessors.

VP evidence showed that the assessors were clearly identifying specific features from the assessment criteria at similar positions within the portfolio. This suggested that they were drawing on a common understanding of those particular features. In many cases these located features were supported by well aligned common expectations about the quality of the found evidence. In some instances it was possible to find evidence of misaligned expectations between assessors. These were most commonly found around interpretations of what constituted a 'detailed' account, a 'basic' description, and the qualities of a 'good' evaluation.

VP evidence also showed that the assessors' search for evidence was influenced by the structural features of the assessment criteria and that this guided their navigation through the portfolio. The location of evidence appeared to be further facilitated by the organisation of the portfolio, with textual cues, such as well placed headings, 'signposting' the location of particular evidence.

'Consistency of performance' appeared to be a factor noted by assessors. Some mentioned the existence of 'coherence' and this appeared to underline their holistic notion of the candidate's true competence.

The assessors appeared to holistically balance the evidence in a number of common ways. They were found to downplay some aspects of the assessment criteria (eg, explaining the purpose of service provision for disabled people; AO1) whilst valuing other aspects particularly highly (eg, the effects of disability on the service user, AO2; evidence of application or generalisation, AO3/4; care value coverage, AO5/6).

The importance of other experts' evidence on their judgment was clearly evident. There were concerns about the quality of the witness statement within the portfolio (which included a third party testimonial about the student's ability to complete an applied task outside of the classroom), and there was a difference in the extent to which this evidence was balanced against the candidate's other performance evidence.

There was also some discrepancy in the way that the evidence of planning and evaluation of the practical task (AO5/6) were dealt with. Some of the assessors appeared to be unsure about the guidance relating to whether these features should be assessed separately or together, in one case affecting the overall judgment of the portfolio.

VP data: Non-AO specific features

Analysis suggested the presence of a number of linked concepts that did not exist in the assessment criteria and to which nearly all assessors alluded to across all but one AO. It was quite common to find assessors commenting on the degree to which the student's work was applied or generic or how they had synthesised information. Although not stipulated in the grading criteria, two of the assessors used synthesis as an indicator of quality. The scale of assessors' attention to application, generality and synthesis (accounting for 20% of the non assessment criteria specific codes) suggests that these overlapping concepts could represent core features for assessors in this area.

It was also possible to find assessors attending to some reference points outside of the portfolio which helped them to arrive at their assessment judgments. Two assessors commented across four different AOs about how the standard of work in the portfolio related to other expectations in parallel qualifications.

Another important feature was perceptions about the influence of the candidate's tutor. Three assessors were responsible for eight comments about aspects of tutor practice. Interestingly, only one of the eight comments could be interpreted in a positive vein. This balance suggests that the assessors were tending to express comments about tutor practice and to explain why they might need to be lenient in their judgments in order to compensate for poorer aspects of the student's overall performance.

Two assessors also commented on how it was important that assessors could interpret the work as being the student's own material. Ten comments on this theme stretched across three different AOs.

KRG data

Assessors elicited 131 constructs over the six Unit 1 AOs. The most senior moderator 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 ($t = 8.16$, $p < 0.001$).

Constructs that were identified by at least three assessors within an AO were interpreted as common constructs. Assessors were also asked to rank each construct in relation to each other, with '1' signifying the most important construct. Mean construct weighting was calculated by totalling each construct weight and dividing this by the number of assessors. Table 3 identifies the four different 'core' constructs which appeared across at least three different AOs. Of these constructs, application is notable because assessors consistently weighted it very highly, suggesting that it might be a very strong core feature of assessment in this area.

Table 3: Common constructs across AOs and their weighting

	Mean weighting (1 = high)			
	1 - 1.9	2 - 2.9	3 - 3.9	4 - 4.9
Application (AO: 1; 2; 3; 5)	3	1		
Quality of description (AO: 1; 2; 3; 5)		2	1	1
Sources (AO: 1; 3; 4; 5)		1	2	1
Example use (AO: 2; 3; 4)		1	2	
Total	3	5	5	2

Within each of the AOs there was a good deal of agreement between the assessors about the qualities of each of the commonly identified constructs. The *application* construct was elicited across four AOs; in two AOs all assessors elicited the construct independently whilst in the other AOs all but one assessor elicited the construct. The *quality of description* construct was elicited across four AOs; in three AOs all assessors elicited the construct independently whilst in the other AO all but one assessor elicited the construct. The *sources* construct was elicited independently by all assessors across all four AOs. The *example use* construct was elicited across three AOs; in one AO all assessors elicited the construct independently whilst in the other two AOs all but one assessor elicited the construct. This is an important finding, suggesting high levels of common understanding for such constructs.

Where agreement levels were lower, careful qualitative analysis of the common constructs data did identify some potentially problematic linguistic issues. These generally appeared to cluster around the issue of unravelling the notions of quality and quantity. Within the *application*; *description or account quality* and *example use* constructs it is possible to find instances where assessors perceived that the concepts of quality and quantity were fused as they progressed through the grade descriptors. In some cases descriptors use adjectives relating to the quality of a concept (eg, simple or basic) alongside adjectives relating to their quantity or existence (eg, some).

The importance of assessors' common understanding of key terminology was apparent in relation to the assessment of *source use*. One key discriminator for assessors judging this concept was their ability to acknowledge the difference between candidates using 'a range of different sources' as opposed to them simply using 'different sources' (AO1, AO3, AO4, AO5, and AO6). The consistent application of this descriptor relies on a consistent understanding of the term 'range'. This concept also had the potential to be interpreted

differently because sources are not always stipulated explicitly within the criteria. This leaves assessors space to infer the appropriate degree of source use to expect at each level.

Finally, for some assessors aspects of qualities expressed in descriptor terminology were perceived to lack 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 (AO2 pass) might also involve it being clear, accurate and detailed (AO2 merit). Similarly, assessors might expect a 'basic' understanding of an issue to be also 'sound' (AO3).

Issues affecting the consistency of judgments

Four key values were identified in the observation and interview data which appeared to influence assessor practice.

A sympathetic and contextualised view of the whole learner

Assessors throughout the system possessed very clear views about the nature of the learners with whom they worked and at whom the qualification was aimed. This element was perhaps reinforced by the fact that all of the assessors, although performing at different levels within the system, retained close contact with learners through teaching commitments.

Many assessors alluded to a dominant learner image that contrasted with 'academic' forms of learning. Assessors talked about the students in their own institutions in terms of them 'not being academic' (M3), typically lacking in self belief (M2), and wanting their achievements to be recognised (M4). This was also reflected in some of the qualification's promotional literature, for example, 'if you find studying boring or difficult, and don't think you'll do well at exams, OCR *Nationals* are the way forward' (OCR, 2007). Furthermore, the need to motivate learners appeared to be a strong core value of the assessors within this system. Assessors commented on the inherent potency of applied learning whilst also talking at length about how the qualification assessment model suited the learners. These features included its lack of testing, its holistic method for judging performance, and a positive assessment approach to recognising students' achievements.

Respect for supportive and positive relationships

Another feature that appeared central to the practice of all assessors was the value that they placed on building supportive and positive relationships with other assessors. This factor was perhaps most clearly evident in the observed interactions between moderators and tutors during school and college moderation visits. The moderators in the study were adamant that they had to focus primarily on making sure that their moderation decisions reinforced the standard expected for accreditation since the risk of not doing would significantly undermine the qualification. The moderators also gave a clear sense that they considered the interactions during moderation visits in a strategic manner. Their concerns

about how to help tutors to improve their assessment practice was an important part of their visits, with moderators structuring their meetings around the times that they could physically feedback directly with tutors and answer their specific questions. This function appeared to be as important as the actual need to moderate and accredit tutors' judgments.

The responsive and supportive tone of the support for assessors also carries potential problems. Assessors 1 and 2 both suggested that an important part of their moderator role was to build good relationships with schools/colleges (with this help being based on their own practitioner knowledge of having taught the qualification). During their school/college visits moderators sometimes needed to choose which aspects of the school or college practice to address since they feared that sometimes the school/college might not be able to deal with all of the possible concerns in one go. Assessors 1 and 2 also suggested that schools or colleges which were new to the qualification might need to be moderated carefully so as not to become overwhelmed with issues. This is a clear concern for a qualification that continues to grow since it presents the obvious potential for moderators to allow for things in some schools/colleges that wouldn't be dealt with in the same way in others.

Valuing professional trust

Some of the assessors felt that an important feature of the assessment model was that it had an element of professional trust built into it, with one suggesting that it contained 'space' for trusting others' professional judgments (T5). The coursework assessment model places the responsibility for assessment decisions onto the tutors. To some extent this reflects the vocational nature of the qualification, where the practices of respected 'experts' have been traditionally connected with their ability to recognise and accredit the practices of others. Again, this implicit value carries potential threats to consistency within the system. One of the tutors expressed her frustration that the qualification might be taught by tutors from non-vocational backgrounds. This suggests that they might lack the intrinsic values that this tutor feels underpins her practice. In one of the observed moderation meetings there was concern expressed that some tutors with a limited understanding of the domain were assessing students to a different standard because their expectations were different.

A commitment to care

The interview evidence showed that the assessors were involved in a variety of 'care-related' activities which didn't directly link to their teaching responsibilities. One of the consequences of this is that they had access to a network of people and ideas that could feed into their practice. All of the assessors had taught other care-related courses in the past, with many teaching these concurrently. This has potentially important implications because it allows some assessors to access some additional tools. Some of the assessors also had strong connections with each other through working on these other qualifications. This meant that there existed a reference network for some assessors within which values might be shared and reinforced.

Five of the assessors also had experience of work in the voluntary care sector, for example working in homeless centres and care homes or through providing drugs and sex education counselling for young people. One assessor was also continuing with her own learning, completing a counselling course, alongside her teaching commitments. This extra-curricular activity enabled some tutors to utilise links established beyond the qualification to augment their students' learning experience. In some cases tutors were able to invite speakers to talk to their students, serving to 'flesh out' the wider context of the qualification, for example, by bringing in current practitioners from the social work sector or through links with other private training providers.

Discussion

This paper attempts to complement the cognitively based VP and KRG data with socio-contextual data in order to shed further light on issues of consistency in a vocationally-related assessment exercise. Activity Theory suggests that assessors who work in different environments might be expected to have differing perspectives.

There might be important methodological concerns about the adequacy of the VP method for understanding expert decision making. In this study levels of verbalisation differed significantly between assessors. Eraut (2000) would suggest that this is unsurprising since some people 'tell more' than others at a similar level of competence. He also points out that there are some kinds of knowledge that are easier to communicate than others and that personal characteristics might interact with this. For these reasons it is difficult to discern whether the VP data really helps to explain variations between different assessors' judgments since we cannot infer from the data whether some assessors were actually attending to more or less of the performance evidence than were other assessors.

Although the methods used for gathering socio-contextual data in this study are partial, and would only be expected to offer a limited insight into any differing perspectives, it is quite noticeable that the assessors in this group shared many key values. This is also mirrored in the data gathered in the parts of this study that have a more cognitive focus. A key finding that emerges within the cognitive data analyses is the particular importance of shared constructs that potentially influence the focus of assessors' judgments. The analysis of supplementary data about assessors helps to contextualise their perspectives and helps to elucidate why particular values are commonly held and sometimes carry more weight.

The importance of illuminating the values that might underpin and link assessor practices is important because it allows consideration of both the positive and negative consequences that might relate to these values. Whilst it might be argued that possessing common values might be a crucial cohesive factor for a community of assessors, supporting consistency of judgment between its members, it might also be the case that these common values might mask contradictory elements. The Activity Theory model that informs this study suggests that incongruous elements exist within a system. Taking a closer look at some of the shared values within a community could lead to a better understanding of where contradictory elements exist and why assessment outcomes might

differ. One of the interesting practices that appears common to the assessors in this group is that of a positive assessment culture which seeks to highlight the achievement of the learner. This contrasts with some of those in other areas of general assessment observed by Sanderson (2001). It might be argued that this practice needs to be seen in the context of Sanderson's 'outer frame' of identified key values. This is important because these values might possess an inherent latent potential to create dissonance within the system and thereby reduce consistency.

An interesting area of tension appears to be the potential for conflict between the assessors' strong philosophical attachment to holistic assessment and the cognitive structures inherent to reading comprehension. The act of reading to judge involves the ongoing integration of evidence through a hermeneutic iteration of past evidence being considered in light of new information which leads to a mental representation of the text (Johnson-Laird, 1983). This was evident in the way that assessors mentioned aspects of different elements whilst explicitly searching for further evidence. This has implications for atomistic assessment models involving heavy written textual documents since it appears that assessors might find it difficult to focus on particular elements in isolation when reading through work.

Another interesting issue raised by this method is the existence of networks beyond the bounds of this qualification that might have had an effect on assessor practice. The portfolio re-assessment exercise found differing levels of inter-assessor agreement, Assessors 1 and 3 exhibited the highest levels and also appeared to have a number of shared frameworks which didn't necessarily overlap with others; these being, an understanding that evaluation requires justification, 'synthesis' being a key quality indicator, and the use of a linear rather than a holistic method when accumulating different elements into a final judgment. It is 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 simply one of the networks (and related tools) that exists to which only some assessors have access.

Conclusions

The benefit of employing an Activity Theory model similar to that of Engeström (1987) throughout this study has been the way that it has encouraged the use of mixed quantitative and qualitative methods to explore the perspectives of assessors. Observing people in their working environment presents the best opportunity to engage with potentially tacit knowledge structures and to widen discourse about assessors' embodied value systems. This also allows a much fuller consideration of the complex factors that can influence assessor practice.

Although theory suggests that the thinking structures of experts are highly tacit and can defy codification this project has elicited evidence using a variety of methods to help to make sense of how some expert assessors work. Whilst the work of Sadler (1989) and

Eraut (2000) might suggest that data should only be interpreted in a tentative way, this project suggests that there is some merit in research studies which use both cognitive and qualitative methods to investigate issues of assessor practice and consistency. This study also suggests that such explorations need to consider the value structures that inform assessors' actions and the ways that they interpret and combine features of assessment criteria.

It appears that assessors' shared focus on care values and applied practice resonates with their personally held value structures. These are evident in the extent to which the assessors tended to engage in active extra curricular care work. This also allowed assessors to supplement the learning experience of their learners by building strong links between theory and practice that further contributed to their students' motivation. The desire to motivate learners also appeared to underpin the positive assessment practices of the assessors. This seemed to reflect the wider culture of the UK vocationally-related learning sector, which is populated with learners who have chosen not to engage with the more 'academic' offers that coexist alongside the *Nationals*. 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. There are also concomitant pressures on the workloads of moderators. On the one hand they are under pressure to complete the moderation paperwork during their school/college visit as well as to foster and maintain good links with schools/colleges that 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.

The great value placed on expert witness testimony is a feature of many vocationally-related areas. Again, assessors felt that this was an important element of the *Nationals*. Shay (2005) has suggested that assessors are often involved in an iterative 'double reading' process where they simultaneously 'read from the outside', utilising the 'official' classificatory schemes for assessment, whilst 'reading from the inside', involving their professional judgment. The use of witness statements might represent such a case. Assessors appear to inherently respect the need to have a competent professional to judge competence within a contextualised learning environment. This carries potential problems where the degree of information provided or its presentation format is inadequate. In such cases assessors' judgments require them to balance their values against the formalised requirements of the qualification. This also resonates with the findings of Greatorex (2005) who found evidence to suggest that witness statements might not support consistent assessment across assessors.

References

Baird, J., Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education Principles, Policy and Practice, 11*(3), 331-348.

Beckett, D & Hager, P. (2002). *Life, work and learning: Practice in postmodernity*. London: Routledge.

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), 943-961.

Curtis, D. D. (2004). The assessment of generic skills. In J. Gibb (Ed) *Generic skills in vocational education and training: Research findings*, pp.136-56. Adelaide, Australia: National Centre for Vocational Education Research.

Dunn, L., Parry, S. & Morgan, C. (2002). Seeking quality in criterion-referenced assessment. Paper presented at the EARLI Learning Communities and Assessment Cultures Conference, University of Northumbria.

Einhorn, H. J. (2000). Expert judgment: Some necessary conditions and an example. In T. Connelly, H. R. Arkes & K. R. Hammond (Eds). *Judgment and decision making: An interdisciplinary reader* (2nd ed.), (pp.324-335). Cambridge: Cambridge University Press.

Elander, J. & Hardman, D. (2002). An application of judgment 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.

Engeström, Y. (1987). *Learning by Expanding: an activity-theoretical approach to developmental research.* Helsinki: Orienta-Konsultit.

Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, *70*, 113–136.

Eraut, M., Steadman, S., Trill, J. & Parkes, J. (1996). *The Assessment of NVQs. Research Report Number 4*. University of Sussex: Brighton.

Greatorex, J. (2005). Assessing the evidence: Different types of NVQ evidence and their impact on reliability and fairness. *Journal of Vocational Education & Training, 57*(2), 149-64.

Hager, P. (2004). The competence affair, or why vocational education and training urgently needs a new understanding of learning. *Journal of Vocational Education and Training, 56*(3), 409-34.

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

Johnson, M. (2007). Is passing just enough? Some issues to consider in grading competence-based qualifications. *Research Matters: A Cambridge Assessment Publication, 3*, 27-30.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R. & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2*(2), 117-146.

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

Kaptelinin, V., Nardi, B. & Macaulay, C. (1999). The activity checklist: A tool for representing the 'space' of context. *Interactions*, *6*(4), 27-39.

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.

Milanovic, M., Saville, N. & Shuhong, S. (1996). *A study of decision-making behaviour of composition markers.* In M. Milanovic & N. Saville (Eds.) *Studies in Language Testing 3* ,(pp.92-114). Cambridge: Cambridge University Press.

Murphy, R., Burke, P., Content, S., Frearson, M., Gillespie, J., Hadfield, M., Rainbow, R., Wallis, J. & Wilmut, J. (1995). *The Reliability of Assessment of NVQs.* Report to the National Council for Vocational Qualifications, School of Education, University of Nottingham: Nottingham.

Nasir, N. S. & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research, 76*(4) 449-476.

OCR. (2007). *'Don't leave your future to chance'*. [Viewed 2 Sept 2007] http://www.ocrnationals.com/learners.html

OCR. (2006). OCR *Nationals* centre handbook. Cambridge: OCR.

Price, M. (2005). Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education, 30*(3) 215-230.

Pope, C. & Mays, N. (1995). Qualitative research: Reaching the parts that other methods cannot reach: An introduction to qualitative methods in health and health services research. *British Medical Journal, 311*, 42-45.

Rapport, F., Wainwright, P. & Elwyn, G. (2005). 'Of the edgelands': Broadening the scope of qualitative methodology. *Medical Humanities*, *31*, 37-42.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Sanderson, P. (2001). *Language and differentiation in Examining at A Level.* Unpublished doctoral dissertation. University of Leeds.

Saunders, N. K. & Davis, S. M. (1998). The use of assessment criteria to ensure consistency of marking: Some implications for good practice. *Quality Assurance in Education, 6*(3), 162-171.

Shay, S, (2005) The assessment of complex tasks: A double reading. *Studies in Higher education, 30*(6), 663-679.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, *39*(2), 229-33.

Stasz, C. & Wright, S. (2004). *Emerging policy for vocational learning in England: Will it lead to a better system?* London: Learning and Skills Research Centre.

Suto, W. M. I. & Greatorex, J. (2006). What do GCSE examiners think of 'thinking aloud'? Interesting findings from a preliminary study. Paper presented at the British Educational Research Association Annual Conference, University of Warwick.

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

Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes.* Cambridge: Harvard University Press.

Watkins-Goffman, L. (2006). *Understanding Cultural Narratives: Exploring Identity and the Multicultural Experience.* Ann Arbor, MI: University of Michigan Press.

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

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

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.

Martin Johnson is a researcher at Cambridge Assessment (University of Cambridge Local Examinations Syndicate). His areas of interest include the impact of assessment mode on performance and behaviour, learners' perceptions of assessment materials, the social implications of assessment, and influences on motivation. Martin has a particular interest in these issues related to vocational and younger learners.
Email: martin.Johnson@cambridgeassessment.org.uk