How can NVQ assessors' judgements be standardised?

A paper presented at the British Educational Research Association Conference, 11-13 September 2003 at Heriot-Watt University Edinburgh.

Jackie Greatorex and Mark Shannon

University of Cambridge Local Examinations Syndicate

Disclaimer
The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Note
This research is based on work undertaken by the University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations (OCR).

Contact details
Jackie Greatorex
Research and Evaluation Division, University of Cambridge Local Examinations Syndicate,
1 Hills Road, Cambridge, CB1 2EU.
☎ 01223 553835
FAX: 01223 552700
💻 **greatorex.j@ucles.org.uk** **www.ucles-red.cam.ac.uk**

How can NVQ assessors' judgements be standardised?

Abstract

NVQ assessors make judgements about the competence of candidates by reference to written standards. They base their judgements on a variety of evidence that is accumulated 'on the job' e.g. observing the candidate working, questioning them and/or judging a product made by the candidates. The tasks that each candidate performs and the other evidence they provide to prove that they are competent is different for each assessment judgement. For NVQ assessment to be reliable each assessors' judgements must be consistent for various candidates and tasks and consistent with the judgements of other assessors. In this situation the question that arises is 'How can NVQ assessors' judgements be standardised (made consistent and reliable)?'

This question is considered by drawing from the research literature and a research study. In the research project a survey of NVQ centres was undertaken to identify what NVQ centres do to standardise assessment judgements. This was followed up by interviews with assessors and candidates from case study centres to consider the issues in greater depth and confirm or deny the results of the survey. After the interviews the assessors took part in standardisation exercises designed and co-ordinated by the Awarding Body. The assessment judgements made in these exercises were made about evidence borrowed from candidates who had already been assessed. It should be noted that the decisions made in the standardisation exercises did not effect the assessment of the evidence that was borrowed for research purposes.

It was found that assessors tended to believe that standardisation was undertaken by ensuring that all assessors followed the same assessment procedure and that such standardisation ensured that assessors made reliable judgements. During the standardisation exercises the assessors found that their judgements were not necessarily always consistent, questioning their belief. The limitations of the standardisation exercises plus the results and implications from the case study research will be discussed.

1 Introduction

This paper is the second in a series about standardising the assessment judgements of National Vocational Qualification (NVQ) assessors. The first paper (Greatorex, 2002) reported the outcomes of a questionnaire survey of 136 OCR NVQ centres. It was found that some centres were undertaking standardisation and that the methods used included discussion, feedback, training, identifying best and/or bad practice. Some of these methods fit with the literature and have been found to facilitate reliability in the context of examinations. However it remains to be established whether these practices actually affect the consistency of NVQ assessment judgements.

The survey was followed by case studies of two centres, these included interviews with assessors and candidates and standardisation exercises undertaken in centre visits. The unit used as a focus for the study was Retail Operations Level 2 Scheme Code 426, unit 2 – 'Meeting customers' needs for information and advice'. The two case studies are reported below, but given the limited size of the research project they are used along with research literature to discuss issues relating to standardisation and reliability.

Standardisation is used here to refer to a process that aims to ensure that:

- each assessor consistently makes valid decisions;
- all assessors make the same decision on the same evidence base;
- all candidates are assessed fairly.

This definition is adapted from the *Joint Awarding Body Guidance on Internal Verification of NVQs* (Joint Awarding Body, 2001). It suggests that standardisation is closely related to two aspects of reliability: intra-rater reliability and inter-rater reliability. Intra-rater reliability is the consistency of assessment decisions made by an individual assessor. Inter-rater reliability is the consistency of assessment decisions (or agreement) between assessors. One of the primary aims of standardisation is, where possible, to improve reliability. Indeed when valid decisions are made with a tolerable level of reliability, assessment judgements could be said to have been standardised. However, the term 'standardised' might also be used to mean 'having been through the standardisation process', even when a tolerable level of reliability has not been achieved.

2 Method

Two centres, Seaside College[1] and Urban College were recruited. Seaside College had one internal verifier (IV) and a team of assessors. Urban College had a more atypical structure with a Project Manager who acted as Lead IV coupled with an arrangement of assessors internally verifying one another.

2.1 Interviews with assessors and IVs

A team of four assessors (including the IV) from Seaside College participated. Two of the team were new to the centre, and one was a relatively inexperienced assessor. A team of six assessors from Urban College participated (including a Lead IV and two IVs who worked on separate contracts, each with a different firm). Two of the assessors from Seaside College were male. All other assessors were female.

2.1.1 Procedure

Semi-structured telephone interviews of approximately 30 minutes duration were undertaken at more or less the same time to ensure that the assessors and IVs could not confer. The interview schedules were informed by the results of the survey and research literature. The

[1] Seaside College and Urban College are pseudonyms.

interviewing was divided between the researchers to avoid interviewer bias. Permission was gained before the interviews were tape recorded and field notes were taken.

2.1.2 Analysis

The tapes were transcribed and qualitatively analysed using a coding frame based on the questions and answers. The interviews were analysed by one of the researchers to ensure that the codes were used consistently.

2.2 Interviews with candidates

2.2.1 Sample

Candidates who had just completed or nearly completed Retail Operations Level 2 were interviewed. Contact was made with the candidates through the centres. There were four candidates from Seaside College and six candidates from Urban College. The candidates from Seaside College were all from the same store, but had different assessors. The Urban College candidates were all from the same company and had the same assessor, but were spread over four stores in the same county.

2.2.2 Procedure

Short face-to-face interviews were undertaken and tape recorded (when permission was given), field notes were also made. One researcher was the interviewer for all the candidate interviews, which took place in the candidates' store(s).

2.2.3 Analysis

A coding frame for the qualitative analysis was developed from the questions and the interview schedule. The researcher who did not interview the candidates analysed the candidates' responses.

2.3 Centre visits

Prior to the centre visit, the assessors were asked to make judgements about four portfolios for the unit from candidates that they did not know. This exercise was repeated after the centre visits, using a different set of portfolios. The assessors were paid for this work. Standardisation exercises, detailed in Appendix 1, were undertaken at each centre after the assessor and IV interviews. These exercises were developed in collaboration with OCR staff who were involved in the *Joint Awarding Body Guidance* development project. The standardisation exercises were based upon research literature about factors which affect reliability in other contexts, and on the experience of the OCR staff. These staff also led the standardisation exercises. After the standardisation exercises, assessors were asked to make judgements about a further four portfolios. The portfolios were borrowed from candidates that the assessors did not know and the portfolios were anonymised. The assessment judgements made in the research did not affect the assessment of these portfolios. When the participants recorded their assessment decisions they also gave reasons for their decisions. The portfolios which assessors judged before and after the standardisation exercises where matched to ensure that there was the same range of ability in each set. The participants completed evaluation forms about the standardisation exercises.

3 Results and Discussion

This section uses the results of the research reported above along with other literature to discuss issues pertaining to standardisation and reliability.

3.1 Current standardisation practice

During the research, participants tended to refer to standardising the **assessment process**, rather than **assessment decisions** per se. It appeared to be that standards were discussed in relation to candidates' performance and assessment decisions were discussed in general, rather than assessors all judging the same portfolio and/or performance and then comparing their decisions and discussing them in relation to the standards and the candidate's performance. The confusion between standardising process and decisions might be because only recently have centres been encouraged to standardise decisions separately from the assessment process. Additionally, it is difficult to entirely separate assessment decisions as they are part of the assessment process. The internal and external verification processes are designed to include checks of the reliability of assessment decisions. This also explains why the *Joint Awarding Body Guidance* project found that little standardisation was undertaken. The answer to the question 'why don't centres standardise assessment decisions?' is that they think that they are standardising them by standardising the assessment process and operationalising the internal verification system. Indeed centres are unaware that they are not standardising assessment decisions.

The belief that standardising procedures, paperwork and/or practice will standardise assessment decisions needs to be challenged. It is likely that standardising processes, procedures and practice will make the NVQ system fairer if assessors judge in a similar way, but it will not necessarily ensure that consistent assessment decisions are made. This is why the standardisation of assessment decisions is crucial. In addition to challenging this belief, the standardisation activities in this research and the *Joint Awarding Body Guidance* need to be promoted, and EVs need to check that these activities are happening and that they are effective. There is also room to improve the standardisation activities used in this study and to adapt them for different situations.

3.2 Traditional assessments and concepts of reliability and validity

3.2.1 Reliability

Traditional assessments, i.e. standardised tests and examinations attempt to be fair to all candidates by requiring everyone to do the same tasks (questions) under the same conditions at the same time. The examination scripts are marked by examiners who do not know the candidates personally. The examiners have all been trained in how to use the mark scheme and their marking is checked and can be subject to scaling.

Traditional concepts of reliability from educational measurement and psychometrics include investigating a test's internal consistency (to ensure that all items are measuring the same trait), inter-rater reliability and intra-rater reliability. Given the arrangements for standard tests/examinations, when item level data has been collected it is relatively easy to statistically measure reliability. For example, the marks awarded by examiners can be compared with those of a more senior examiner.

3.2.2 Validity

"Conventional approaches to validity, or whether we are measuring the things we want to, are usually concerned to do one of two things. One approach is to look at patterns of relationships between the measures we have, and sometimes others which may be available

elsewhere, to try to establish whether the results we have obtained are theoretically or logically justified (construct validity). Alternatively we can see how well measures predict subsequent events such as aspects of performance at work (predictive validity). A further possibility is to see how far measures relate to contemporary events (concurrent validity)" (Massey and Newbould, 1986, 96).

3.3 Inappropriateness of traditional concepts of reliability and validity to NVQs

3.3.1 Reliability

Statistical measures of internal consistency cannot be meaningfully calculated for NVQs. Measures of internal consistency assume that the scores from individual items can be added up to give a final score. NVQ performance criteria cannot be treated as individual items for which candidates score 1 or 0 and then add together to make a score. This is because, as Wolf (1995) reports, assessors make holistic decisions about candidates, they do not make individual decisions about each performance criteria and then 'add them up'. The holistic approach is contrary to the NVQ specifications, which demand that all evidence requirements, knowledge requirements and performance criteria (PCs) must be met. Typically there are over a thousand separate assessment judgements to be made throughout the process of showing and determining competency for an NVQ (Eraut *et al*., 1996), so following the strict NVQ line might be unrealistic.

In the NVQ context each candidate undertakes different tasks (which ideally arise as naturally occurring evidence) which is used to illustrate their competence. In this scenario inter-rater reliability is a concept of limited value. However in some circumstances, for example, an internal verifier checking the assessment judgements of an assessor or in a standardisation exercise, a number of assessors could all assess the same candidate doing the same task. In our study, and in research by Murphy *et al* (1995), a number of assessors all made judgements about the same portfolios. This approach is limited, as if none of the assessors know the candidate, and cannot question the candidate, the situation is unrealistic. If one assessor is more familiar with the candidate they could provide background information but then the assessments are tainted by the thoughts of the assessor who is familiar with the candidate. Alternatively two (or more) assessors could observe the candidate, an approach which was used by Eraut *et al* (1996) to investigate reliability. However it might be off-putting for candidates to be observed by two or more observers. Two of the ten candidates interviewed in our research had experienced their assessors being observed and they did not say that this was problematic. Intra-rater reliability could be investigated using similar techniques, but it too is a limited concept in the NVQ context for the reasons given above.

There's a paradox: internal assessment by an assessor who knows the candidate is meant to be more valid than external measurement exactly because that assessor has extra knowledge, yet this extra knowledge must reduce any measure of inter-rater reliability.

3.3.2 Validity

It can be argued that conventional examinations and tests are of limited validity if they do not meet the goals of education and training (Jessup, 1991). Traditionally, General Certificate of Education (GCE) A levels have been for the academically orientated young people who will go on to university, and vocational education (training for jobs) has been reserved for a different group of young people (Wolf, 1995). Clearly these two different types of education serve different purposes. Consequently GCEs were traditionally not a valid indicator of a prospective job seeker's skills. The validity of conventional examinations and tests is also questionable on the grounds they are taken at the end of a course rather than as part of a learning programme and they do not provide naturally occurring evidence of competence (Jessup, 1991). This last point is partly why NVQs were introduced as a competence based criterion referenced assessment, where evidence for competence was derived from the workplace.

In the context of NVQs, Jessup (1991, 192) said: "Validity is concerned with the extent to which an assessment instrument or method actually measures what it is designed to measure. It implies comparison between the assessments and some external criterion, i.e. that which one is trying to access. Within the NVQ model of assessment one clearly has an external reference point for assessment - the statement of competence." Eraut and Steadman (1998) defined the external validity of a competence based award as the match between the competence measured by the award and the performance of people deemed competent in the workplace. Such an approach to validity encompasses construct validity and concurrent validity.

It was Jessup's view that: "…validity is the objective of assessment, while reliability is only important in so far as it contributes to valid assessment decisions." Jessup (1991, 50). From this perspective reliability (and by implication standardisation) is of relatively limited importance.

Wolf (1995) reports that proponents of NVQs maintained that detailed and specific criteria guarantee reliable assessment. However there have still been questions raised about levels of reliability and the importance of reliability. For example, the first independent study of NVQs was commissioned as part of the Beaumont report (a government commissioned review of NVQs) and was conducted by Murphy *et al* (1995). Later the issue of reliability has re-emerged in the emphasis that the *Joint Awarding Body Guidance* places on standardisation (and therefore reliability).

3.4 Are NVQs valid?

NVQ standards were developed using functional analysis (Mitchell, 1995) This method focuses on occupational roles and the expectations of people performing the role. Functional analysis involves interviewing practitioners. The process begins by clarifying the key purpose of the occupation, followed by a sequential analysis of the details of the required activities. The summary phrase (description of the key purpose of the occupation) is refined by asking practitioners pertinent questions to develop a list of the tasks and roles that are needed to fulfil the occupational purpose. The purposes, tasks and actions are organised into a functional map indicating the relationships between them. The constituent details can be used as performance criteria. Functional analysis is used to develop a profile of an occupation rather than a qualification. The functional analysis does not rank the standards in importance or in a hierarchy (Mitchell, 1995).

When evidence is not naturally occurring the validity of NVQs might be compromised. The centre visit to Seaside College and interviews with candidates (from both colleges) revealed that the evidence that they used in their portfolios was not all naturally occurring; the candidates completed a workbook to meet some of the NVQ standards. Earlier Grugulis (2000) reported a similar result: she found that the majority of candidates registered for a management NVQ at three case study organisations were spending a great deal of time collecting evidence and undertaking tasks for their NVQ which were not really part of their job.

3.5 Are NVQs reliable?

When NVQs were first introduced, it was argued that reliability was secured through specific written PCs. In our research the assessors (and IV) from Seaside College all said that there was an agreed interpretation of the standards at their centre. The Lead IV from Urban College and all the assessors except one, thought that there was an agreed interpretation of the standards. But at both colleges there were a diversity of ways of interpreting the part of the standards which says 'observed on three occasions'. The IV at Seaside College, and two members of the team from Urban College, said that individuality in interpretation of the standards could be tolerated. The IV at Seaside College added that individuality was acceptable as long as the team were working to national standards. One assessor from

Urban College said that if one assessor was more lenient or severe than others it was not fair to candidates.

Standardisation meetings attended by the authors illustrate that there is some disagreement between NVQ assessors at the PC level and sometimes at the unit level (see Table 1). Ten assessors made judgements about the same sixteen portfolios which they assessed using the same NVQ Retail assessment criteria. For all portfolios there was an agreement of 50% or more of the assessors that the candidates were not yet competent. When assessors disagreed about the competence of a candidate, they tended to give varied reasons for their different judgements. It could be that this is the individuality in interpretation that the assessors mentioned.

The candidates from Seaside College all thought that an assessor would apply standards in the same way, with different candidates. But one candidate from Urban College thought that standards might be applied differently to different candidates. She suggested that an assessor might assume that a senior worker would be more competent and so might be prepared to base an assessment decision on less evidence. The candidates from Seaside College said that different assessors would apply standards in the same way. Generally, the candidates from Urban College were confident that their assessor applied the standards rigorously, but they were not confident that other assessors would do so. Two of the candidates from Urban College had only had one assessor so they did not feel that they could realistically comment about this issue.

Despite the difficulties involved in measuring NVQ reliability, researchers (including the authors) have investigated the consistency of NVQ judgements. Eraut *et al*. (1996) visited six contrasting assessment sites. They observed assessors while they were assessing candidates, interviewed seven assessors, seventeen centre managers and a number of trainees, and gathered documentary evidence. They also received questionnaire responses from 1,233 assessors. One of their findings was that assessors from the same centre generally agreed with one another about assessment decisions while they were being observed by researchers. They suggested that the assessors did not feel that they could disagree in front of the researchers. One of their conclusions was that the claim that the assessment of NVQs was reliable could not be sustained. In other words, writing specific detailed PCs did not guarantee reliability.

Murphy *et al.* (1995) undertook a project in which assessors judged a small number of units in their sector of expertise (Bricklaying, Business Administration, Care, Supervisory Management or Vehicle Maintenance). Eraut *et al*. (1996) discuss an aspect of the Murphy *et al.* study, in which units taken from candidate's portfolios were each judged by between three and seven assessors. Of the thirty-five units assessed, sixteen were judged to be both 'competent' and 'not yet competent' by different assessors. Eraut *et al*. argued that if the assessors were considering a large ability range, this level of disagreement would be more significant or problematic than if a small ability range were being considered. However, it is difficult to establish the width of the achievement range in Murphy *et al*.'s (1995) study. Therefore, it is difficult to judge the true level of disagreement between the assessors.

Given the evidence considered, it could be argued that the NVQ competence approach is not necessarily valid and that it does not guarantee or necessarily lead to reliability. This would be unfair to candidates and we need to look for ways to improve the situation.

3.6 How can NVQ assessment be made fairer?

3.6.1 Communities of practice

If the fairness of NVQ assessment is to be seen to be improved, a number of measures might be taken. One approach is to tackle the problem that the reliability of NVQs might be lower than is desirable. Konrad (1998a & b) argued that the literature about communities of practice could be used to improve the training of NVQ internal verifiers (and consistency in

assessment judgements). Eraut and Steadman (1998) support this view, concluding that the training of assessors and verifiers was not focused sufficiently on building assessment communities that would result in consistent and comparable decisions. In the context of vocational education in Australia, Jones (2003) suggests that a community of assessment practice facilitates a common frame of reference for assessors, and promotes agreement between them. A community of practice is a group or network of people who work together and share similar goals and interests. To meet these goals they use common practices, work with the same tools and use a common language. This shared activity facilitates similar beliefs and value systems e.g. how to interpret NVQ standards.

One move towards a community of assessment practice is the founding of the Institute of Assessors and Internal Verifiers which will be a forum for assessors and verifiers to meet and network. The Regional Offices of Awarding Bodies also hold assessor and verifier network meetings for similar purposes. Results from the interviews in our research show that generally the assessors believed that they were part of a tight knit team (within their centres) although it was acknowledged that assessing was "very much a lone role". There was communication between staff despite the distances between them. The community was facilitated by meetings, mobile phones, contact with IVs and memos.

A community of practice facilitates reliability by establishing a shared agreement on the interpretation of standards. The assessors at the Seaside College and Urban College generally reported that they had a shared understanding of the standards but they did have different interpretations of the phrase 'observed on at least three occasions'. The IV at Seaside College explained that individuality in the understanding of the standards could be accommodated as long as everyone was working to the national standards. After the standardisation meeting at Urban College one assessor commented that they had thought that they had a common understanding of the standards and that they applied them in the same way but that that was not necessarily so. However the levels of agreement amongst the assessors (see Table 1) suggests that there was some commonality between the judgements that the assessors made, even though they might have reached the same decision for different reasons.

3.6.2 Standardisation exercises

Joint Awarding Body Guidance outlines some standardisation activities (see Appendix 3). The standardisation activities which were undertaken as part of our research were generally received positively (see Table 2 and Table 3). The most striking finding was that only one participant had taken part in a similar exercise before (see Table 2). This tallied with the comment from the Lead IV at Urban College that the team had not all assessed the same portfolios and evaluated their decisions against the standards and the candidates performance. (This was essentially the basis of our standardisation exercises). Generally, the standardisation workshops were received positively. The general opinion of the assessors was that the activities were useful for standardising assessment decisions, and that the standardisation meeting would improve their professional practice. This is in sharp contrast to the assessment judgements they made (see Table 1). It seems that the standardisation meeting had a limited effect on the level of agreement between assessors. When considering these results we should be mindful that there were only ten participants in the standardisation meetings. The standardisation exercises used in our research have their limitations, but they are a step in the right direction.

3.6.2.1 Difficulties in following standardisation procedures

It is often impractical to make candidates' performance available to a number of assessors at the same time. Examples of candidates' work might be a conversation with a customer, building a staircase, driving a forklift truck etc. As mentioned earlier, for a number of assessors to be able to assess these they would all need to watch and question the candidate. Assessment and standardisation must be fair to the candidate, who might not appreciate an audience. Only two of the ten Retail candidates interviewed in the UCLES

study had experienced their assessor being observed whilst they were being assessed. So large numbers of assessors observing a candidate together would be unusual for candidates and probably inappropriate. A pragmatic solution is for all assessors to assess the same portfolios, videos/DVDs of performance or role plays or an artefact made by the candidate. Obviously there could be one copy of each portfolio, which is passed around. This is one suggestion made in the *Joint Awarding Body Guidance* (see Appendix 3).

A limitation of the use of copies of portfolios/videos/DVDs in standardisation exercises is that it is not the candidate's performance that is being assessed, but a record of the candidate's performance. Additionally, the candidate is not available for the assessors to supplement their observations with questions about the candidate's performance to assess the candidate's knowledge and understanding. So the assessor is asked to make decisions in an unrealistic situation and with limited evidence (a point made by some of the participants in our research). Hence the validity of the standardisation activities is affected.

Some centres might be disadvantaged if Awarding Bodies demand that exercises such as those suggested above are followed. Smaller centres, such as small shops, might have limited resources, and might not have a time and place to meet. Others might not have the equipment to play a videotape of an activity to be assessed. Assessors from one centre can be spread across a wide area and it can be expensive and time consuming to bring them together for standardisation meetings. This was the situation with the two case study centres in our research.

If a number of assessors make judgements about a candidate's performance, and the candidate is more familiar to some than others, then the assessors will be making decisions, based upon different information. Simplistic comparisons of these decisions might result in inaccurate estimates of levels of reliability. Additionally, the validity of the assessment decisions might be compromised. If the performance of a candidate who was known to a couple of assessors was used in a standardisation exercise, other assessors could be given information about the candidate as an aid to decision making. As mentioned earlier this might make the standardisation exercise more valid, as it would more closely resemble the conditions under which assessment decisions are usually made. However, those taking part in the exercise would have the assessor's perception of the situation, rather than having their own experience to work on. This is likely to facilitate unrealistically high levels of agreement, rather than identifying points of disagreement for the assessors to discuss.

In standardisation exercises at the two centres attended by the authors, disagreements at both the PC and the unit level were discussed. Given that assessors make holistic decisions, and do not make separate decisions at the PC level, perhaps it would be useful to encourage assessors to be more concerned about disagreement at the unit and qualification level. Disagreement at the PC level could be used to facilitate discussion and develop mutual understanding of the different possible interpretations of the PC.

Another problem with trying to offer centres guidance about standardisation exercises is that, as the traditional approaches to reliability do not apply, the researchers are unsure how much disagreement (inconsistency at the unit or PC level) is tolerable. The standardisation meetings resulted in varied levels of agreement, depending upon the portfolio used (see Table 1). For portfolios on the borderline between competent and not yet competent we would have to establish what represents an acceptable level of agreement. For instance would a figure of 75% of assessors reaching the same decision represent a credible level of agreement?

Before the data collection at standardisation meetings we had believed that competent, not yet competent and insufficient evidence were mutually exclusive categories. But some participants wanted to put portfolios in more than one of these categories. A response of competent and insufficient evidence might suggest that on the evidence presented the candidate is likely to be competent but in order to claim the criteria they would need more evidence. A response of not yet competent and insufficient evidence might suggest that there is insufficient evidence of competence or that from the lack of evidence provided it is likely that the candidate is not yet competent and that the evidence is lacking. The overlapping of

categories might be due to the research context. In practice the assessors would have been able to question the candidate to gather more evidence upon which to make their judgement - they had a lack of evidence because they only had the written evidence in the portfolio, they had not come across the candidate or their performance. A similar problem arises when IVs sample candidates portfolios which are sometimes lacking in details that the IV feels they need to make a judgement. Obviously they can overcome the problem by consulting the assessor to clarify issues. For future standardisation exercises and/or research would be advisable to define 'insufficient evidence' carefully; does it refer to 'insufficient evidence to decide' or 'insufficient evidence of competence'?

Examples B and D of the *Joint Awarding Body Guidance* (see Appendix 3) concern the use of evidence in NVQ assessment. The research described above has focused on the PCs, and judgements at the unit level. Standardisation exercises must incorporate evidence requirements as well as PCs, which are applied together. Should researchers be focusing on the levels of agreement about whether evidence meets particular evidence requirements?

The issues above have been tackled from the perspective of trying to undertake standardisation in centres and providing guidance on best practice. The problems outlined above apply equally to measuring reliability of NVQs in research studies.

3.6.3 Other approaches to make NVQs fairer

It has been found in the context of examining that certain factors affect the reliability of assessment. These include feedback to assessors, discussion between assessors about candidates' work in relation to written standards, observations of assessments, and other factors, for a full discussion of these issues see, for example, Wolf (1995) or Baird *et al* (2002). From the interviews with assessors and candidates there was evidence that assessors were observed assessing by IVs, that portfolios were checked by IVs and EVs, that assessment decisions were discussed and that feedback was given to assessors about their decisions. These measures are all part of the verification process.

In our research, some assessors made other suggestions for developing common understandings of the standards. For example, one assessor said "I think it would help if all the Awarding Bodies were working to the same hymn sheet. We seem to get mixed reactions from different EV's and from different Awarding Bodies. I think sometimes, because we work with a lot of different Awarding Bodies like OCR, City & Guilds etc and we'll get a visit from City & Guilds and the EV will tell us one way and the OCR will want it done a different way. Then there is another set of information they want, which is the same information so sometimes it seems a little bit inconsistent. ….. you have to try and remember what different EVs are looking for." It was thought that there would be greater standardisation if centres experienced consistency from EVs and different Awarding Bodies. Awarding Body guidance should be increased and made more specific and constructive. On this matter one assessor said: "I think just giving clear guide lines from the Awarding Bodies about exactly what sort of evidence they want. When they say this number of PC's don't have to be observed well, you know, they always say use simulation/role play, which no one ever uses. They always give you those standard things …but they should give you something constructive they would really like you to use…and you can use.…….All those things that they suggest are never any good, …they say use witness testimonies but we've had so many bad reports on the witness testimonies that we've produced through our EV so now we don't bother using them. Officially a witness testimony should be from the manager recording an occasion when a candidate has done something. Managers never have the time they're always too busy. The candidate writes them and then you go over and try to discuss it with the manager, witness testimonies just doesn't work in the real world."

Many of the comments that the assessors made about the reasons for their decisions were about the evidence upon which those decisions were based. Johnson (1995) found similar results in her study about consistency of judgement in Business Administration Level III. She found that there was some disagreement between assessors about whether candidates were competent or not yet competent and she concluded that: "It would appear that the

discrepancies often occur as to the level of acceptability of the evidence submitted by the student." (Johnson, 1995, 25). This raises questions about whether judgements could be made more reliable by restricting the evidence that is accepted for assessment.

NVQ specifications include evidence requirements as well as PCs. This is an attempt to increase the consistency of assessment judgements by restricting the diversity of evidence that can be used. Research (referenced above) shows that some evidence is not naturally occurring. There is arguably a case for requiring that candidates all undertake the same (or similar) tasks and that the tasks are assessed by external examiners. One way of reaching a common understanding of PCs and improving the reliability of NVQ assessors' judgements would be to restrict the types of evidence that can be used or the tasks that must be assessed. At present, some evidence requirements specify the types of evidence that the candidate can/must use, but they do not restrict the tasks that must be assessed. In standardised tests, and indeed in public examinations, the situation is quite the reverse. All candidates are required to undertake the same tasks (or there are restricted options) and the test/examination is administered to everyone in the same way. For A levels, GCSEs etc there is more flexibility in coursework, where the task is set by the teacher within boundaries outlined by the Awarding Body. The ideal solution might be to specify the task without compromising the validity of the assessment.

Luty (1999) suggests that the tasks that NVQ candidates undertake should be specified, and that they should combine an academic and a competence based approach to assessment. One way to do this is to use work based portfolio assessment (WPA), which is evidence that is presented in the workplace rather than in a paper portfolio. Evidence might include a guided tour of the workplace when the candidate produces evidence from files (paper and electronic) whilst giving a verbal explanation of how the information is used. There are some practical problems with the approach e.g. time, cost and creating an audit trail. Fowler (1999) says that when WPA was trialed, everyday occurrences were used as evidence, and it was well received by stakeholders as it was genuinely work based. In fact, there has been a 240% increase in NVQ management registrations for EdExcel since this method of assessment was introduced. He adds that when this approach to assessment is used the assessors' judgements are more valid, as they are made in the workplace, rather than assessors being tempted to imagine the workplace and assessing a paper based portfolio remotely. Obviously this method of assessment is not the solution to everyone's problems, and some candidates still prefer the paper based portfolio (Fowler, 1999). For this approach to work well assessors would need to be trained in the meaning of performance criteria in the workplace.

Recently Bowen-Clewley (2003) has been promoting the use of professional conversations as a method of assessment along with other methods in the vocational context. Professional conversations are based upon Devereux's (1997) professional discussion as a form of assessment. In a professional conversation the candidate and assessor meet as equals. The candidate orally presents their case to the assessor that they have reached the standards, and refers to naturally occurring evidence as necessary. To prepare for this meeting, the candidate meets with a facilitator to discuss the standards and suitable evidence. The assessor begins by assuming that the candidate has reached the standards but asks questions after the candidate's oral presentation to ensure that the work is theirs and to fill any gaps in information. Such an approach requires that there is a common understanding of standards and evidence requirements between the facilitator and assessor. If the evidence is naturally occurring then this is a valid method of assessment. Bowen-Clewley (2003) reports that this method is economical and cost effective. This method of assessment relies heavily upon the skills of the assessor and Bowen-Clewley (2003) has trained assessors in this method. Given that the new Assessor and Verifier units require that a professional discussion be used as a form of evidence, further investigation of this assessment method might be beneficial.

4 Conclusions

This paper has raised a number of issues, some of which are well rehearsed, and still leaves some loose ends. Given that UCLES is still involved in researching the reliability of NVQs and the standardisation of assessor judgement, it might be that more answers will be available at a future date. The authors welcome suggestions for methods of reaching common understandings of performance criteria and evidence requirements, and for addressing any of the issues raised in the paper.

5 References

Baird, J., Greatorex, J. and Bell J. F. (2002) *What makes marking reliable? Experiments with UK examinations*. International Association of Educational Assessment conference, October 2002, Marco Polo Hotel, Hong Kong

Bowen-Cleweley, E. (2003) *A review and reflection on current practice in assessment and moderation in New Zealand*, Invited lecture given at the University of Cambridge Local Examinations Syndicate, 14th August 2003.

Devereux, C (1997) *Rigour Without Rigidity* WA Consultants

Eraut, M, & Steadman, S. (1998) *Evaluation of Level 5 Management NVQs Final Report 1998. Research Report Number 7*. University of Sussex: Brighton.

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

Fowler, B. (1999) Better or worse? *T magazine*, ww.tmag.co.uk

Greatorex, J. (2002) *Two heads are better than one: Standardising the judgements of National Vocational Qualification assessors*, A paper presented at the British Educational Research Association Conference, 12 to 14 September 2002 at University of Exeter.

Grugulis, I. (2000) The Management NVQ: a critique of the myth of relevance, *Journal of Vocational Education and Training*, 52 (1), 79-99.

Jessup, G. (1991) *Outcomes NVQs and the Emerging Model of Education and Training*, The Falmer Press, London.

Johnson, C. (1995) Achieving consistency in NVQ assessment, In Peter McKenzie, Philip Mitchell and Paul Oliver (eds), *Competence and accountability in education*, p19-28, Aldershot: Arena.

Joint Awarding Body (2001) *Joint Awarding Body Guidance Internal Verification of NVQs*, Joint Awarding Body: London.

Jones, A. (2003) *Judgement calls: How TAFE teachers make assessment judgements*. Unpublished Ed D, Monash, Melbourne.

Konrad, J. (1998a) *Assessment and Verification of National Vocational Qualifications: a European quality perspective*. Education-line www.leeds.ac.uk/educol/index.html.

Konrad, J. (1998b) *The assessment and verification of National Vocational Qualifications [NVQs]: a European quality perspective*. Education-line www.leeds.ac.uk/educol/index.html.

Luty, C. (1999) A Coming of Age? *T magazine*, www.tmag.co.uk

Massey, A. and Newbould, C. (1986) Qualitative records of achievement for school leavers: an alternative approach to technical issues, *Cambridge Journal of Education*, 16 (2), 93-99.

Mitchell, L. (1995) Outcomes and National (Scottish) Vocational Qualifications. In Burke, J. (1995) (ed) *Outcomes, Learning and the Curriculum, Implications for NVQs, GNVQ's and other qualifications.* The Falmer Press: London.

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.

Wolf, A. (1995) *Competence Based Assessment*. Open University Press: Buckingham.

Appendix 1 Standardisation Exercises

Standardisation Activity 1 'Consistency in individual assessment decisions'

The assessors were asked to make comments about the content of the four portfolios that they had just made assessment decisions about and to discuss their assessment decisions. The IV(s) was asked to comment last. The portfolios and associated assessment decisions (including the OCR assessment decisions) were discussed one at a time.

After this discussion the assessors were given another form and asked to record their assessment decisions about the four portfolios again. The purpose of this exercise was to identify whether the assessors changed their minds after discussing their assessment decisions and gaining feedback from the OCR EVs.

Standardisation Activity 2 'Customer in need' – an assessment decision

The OCR personnel, accompanied by a researcher, acted out two scenarios (role plays) devised by one of the OCR staff. Each scenario had to be adapted a little to the environment of the room where it was acted. That is, at one centre the scenarios took place in a stationary shop and in the other at a luggage store.

Scenario 1

Customer enters shop	Approaches counter.
Assistant busy	Looks up, doesn't acknowledge or smile but waits for customer to speak first.
Customer 1	Requests a previous order.
Assistant	Asks for details - SHOUTS to colleague - 'woman here wants to know if order's come'.

No reply. Leaves shop floor to ask and returns with a very long explanation.

Customer 2	Enters, he is ignored.
Customer 1	The time the customer would have to wait for the order to arrive is too long.
Assistant	Alternatives offered which she informs is 'the best on the market' but informs customer she can get same 'down the road at .....'
Customer 1	Mollified but not satisfied - says assistant is improvement on last girl's efforts.
Assistant	Ooh! Do tell - what did she do/say - who was it?
Customer 1	Doesn't want to gossip and gives a curt reply and leaves assistant with customer 2.
Assistant	Well, I know she has a problem with my colleague but she doesn't need to take it out on me!

Scenario 2

Customer 1	Enters and assistant immediately smiles and speaks.
Customer 1	Requests a previous order.
Assistant	Asks for details - using customer name throughout transaction - goes to exit and calls supervisor's name - no reply. Informing the customer what she is doing, she uses telephone to contact supervisor and quietly gives details.
Customer 2	Enters - assistant acknowledges them and then proceeds to explain briefly the order delay - offers alternative.
Customer 1	Has difficulty with delay - alternatives shown by assistant.

Customer 1	Complains about other staff by comparing service.
Assistant	Apologises, and offers a 'complaints' form – informs customer 1 that the details are confidential and only the manager deals with them.

The first scenario was acted out and the assessors were given 'assessment record and feedback sheets' to complete whilst they observed the role play. They recorded their decisions of competent, not yet competent, or insufficient evidence on the RED sheets provided. Then the assessors, IV and OCR personnel discussed the assessment decisions and the performance with relation to the standards. This process was repeated for the second scenario. The scenarios were designed so that one candidate was clearly competent in this scenario and the other was not.

It should be noted that the decisions made in this activity were different from the decisions in the other activities, as the decisions were about simulated performance and not written evidence.

Standardisation Activity 3 Standards - 'Out with the old, in with the new'

There are two sets of live standards for two years from 1 April 2002, and the old standards are used with candidates who registered before this date. The 'new' standards are applied to candidates who registered after this date. The assessors judged two portfolios against the new standards and recorded their decisions. After participants had made their individual assessments, the assessment decisions that the assessors and OCR personnel had made, the contents of the portfolios, and the differences between the old and the new standards were discussed in relation to one another. After the discussion, the assessors recorded their final assessment decisions on the RED forms provided.

Standardisation Activity 4 'Consistency in assessment decisions 2'

The assessors were presented with four portfolios. The assessors read the portfolios for themselves. Then they decided whether they agreed with how the claims that the candidates has made about their competence and whether the candidates were competent or not yet competent or whether there was insufficient evidence. Their assessment decisions and the content of the portfolios were then discussed in relation to the standards. After this discussion, the assessors completed RED forms indicating their assessment decisions.

After the standardisation exercises were completed, the assessors were left with four more portfolios to make individual assessment decisions about, and RED forms to record these decisions.

At the end of the centre visit, the assessors completed an evaluation form indicating whether they had undertaken any standardisation activities like the ones described above, and explaining how useful they found the activities.

Appendix 2 Results from the standardisation activities

Table 1 Frequency with which assessment decisions were made by candidate/portfolio

	Portfolio	Competent	Not yet competent	Insufficient evidence	Missing
Pre-standard-isation judgements	P12	0	9	1	0
	P13	2	6	2	0
	P14	0	9	1	0
	P15	1	6	2	1
	P1	1	9	0	0
	P2	0	10	0	0
	P3	0	9	0	1
	P4	1	7	0	2
Post-standardisation judgements	P9	2	7	1	0
	P20	2	8	0	0
	P21	0	7	3	0
	P22	1	7	2	0
	P6	1	9	0	0
	P7	1	8	1	0
	P23	2	5	3	0
	P24	1	5	4	0

Note: The portfolios P12, P13, P14, P15, P1, P2, P3 and P4 were the portfolios on which pre-standardisation decisions were made and post-standardisation decisions were made on portfolios P9, P20, P21, P22, P6, P7, P23 and P24.

Table 2 Frequencies of responses about the usefulness and prevalence of standardisation activities

Activity	Activity (number) was useful for standardising NVQ assessor decisions					Have you ever undertaken this activity before?	
	Strongly agree	**Agree**	**Disagree**	**Strongly disagree**	**Don't know**	**Yes**	**No**
1	4	6	0	0	0	0	8
2	3	6	1	0	0	0	9
3	3	6	1	0	0	1	7
4	3	6	0	0	0	1	6

Table 3 Evaluation of the centre visit

Question	**Strongly agree**	**Agree**	**Disagree**	**Strongly disagree**	**Don't know**
The aims and objectives of the training day were clearly stated	2	6	1	0	0
The aims and objectives of the training day were achieved	3	6	0	0	0
The training day will effect my professional effectiveness in this area	4	5	0	0	0
The training day was suitably participative	4	4	1	0	0
The contact was as I had expected/hoped	1	6	1	1	0
The trainer(s) was appropriately skilled and informed to deliver the event	4	5	0	0	0
The trainer's style was conducive to learning	4	4	1	0	0
The pre-event administration was satisfactory	2	6	1	0	0

Appendix 3 Extract from Joint Awarding Body Guidance

6.Standardising Assessment Judgements

6.1 The third strand to verifying assessment is to standardise assessment judgements.

Aims:
1. To ensure that each Assessor consistently makes valid decisions
2. To ensure that all Assessors make the same decision on the same evidence base

6.2 Issues/concerns

The project has raised the following issues/concerns about the effectiveness of this aspect of internal verification:-

- negligible use of standardisation exercises within Centres
- poorly conducted standardisation

6.3 Standardisation (sometimes referred to as benchmarking or moderating) is an important part of Internal Verifier duties. In many Centres visited in the project, this aspect of the Internal Verifier role was substantially underdeveloped, often relying on informal contact between Internal Verifiers and their assessment team to ensure a common standard of decision making. And although team meetings were held these were frequently poorly attended and concentrated on relaying information and/or tracking candidate progress. These issues are clearly important but it is critical, **particularly for Centres with a number of dispersed, peripatetic or inexperienced Assessors** that standardisation exercises are undertaken with all the Assessors on a regular basis.

6.4 The simplest means of completing a standardisation review is to collate copies of evidence presented for unit accreditation and ask each Assessor to make a decision based on what is in front of them. It is also helpful to ask them to note any queries they may have e.g. further information needed or authentication of a piece of evidence. This enables the Internal Verifier to check that Assessors are asking the right questions when looking at portfolio evidence as well as arriving at the correct decisions i.e. that the **process as well as the judgement is sound.**

The following examples should provide ideas for Internal Verifiers to carry out such an exercise.

Example A

Select a "problem" unit from a qualification, which many of the team assess and ask each to bring along two examples of completed units they have signed off. The units are then passed around the group and each Assessor completes an assessment feedback form as if they are assessing a unit and providing feedback to a candidate. Discussion follows. Sheets are collected and evaluated by the Internal Verifier and feedback given to individual Assessors, confidentially, at a later date.

Example B

Concentrate at one session on particular types/sources of evidence and how they are assessed, including the recording of the assessment. For example, each Assessor could bring a number of witness testimonies from their candidates', or examine observation records. The group then share constructive criticism about items tabled (which may be made anonymous for the purpose of the exercise).