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

Disclaimer
The opinions expressed in this paper are those of the author 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 data analysed by the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations.

Contact details
Jackie Greatorex, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU.
greatorex.j@ucles.org.uk.

Available at www.ucles-red.cam.ac.uk

Two heads are better than one: Standardising the judgements of National Vocational Qualification assessors

Abstract

There has been considerable research to measure the reliability of National Vocational Qualifications (NVQs) and explore the factors that might influence assessors' judgements. The literature suggests specific factors, for example, training, discussing assessment judgements and candidates' work, that contribute to the improvement of reliability in other areas of assessment. However there is little evidence indicating what methods are used in NVQ centres to standardise assessor judgement. This paper presents results from a questionnaire sent to Retail Operations NVQ centres indicating that centres undertake standardisation. Some of the methods they use tally with those given in the literature about improving reliability. NVQ centres also use methods of standardisation not found in the literature. The advantages and disadvantages of the methods are outlined. Below it is concluded that the results of the questionnaire are generally positive and further research should evaluate the effectiveness of standardisation.

Introduction

National Vocational Qualifications (NVQs) are assessed against criteria to ensure that candidates have reached the required standards. Some of the criteria are linked to occupational standards, although there is no official syllabus for NVQs, there is an official pedagogy; the awards are meant to reward students for taking responsibility for their own learning. NVQs are made up of units. To complete an NVQ a candidate must have met all the criteria for the qualification. Candidates often use a portfolio to store evidence that they have met these assessment criteria. In the NVQ system *Centres* appoint their own assessors and Internal Verifiers for each qualification. The latter verify that assessment decisions are valid. *Assessors* are competent according to specified occupational standards in their area of expertise. All assessors, whatever their area of expertise, should hold the Units D32 (Assess candidate performance) and D33 (Assess candidates using different sources of evidence). These units are part of the NVQ framework. An assessor cannot 'sign off' units for a candidate unless the assessor has the appropriate 'D' units. They judge candidates' performance and knowledge evidence against the NVQ standards, as well as provide constructive feedback to their candidates and sign off candidates' completion of their NVQ. They also maintain the level of their professional competence (Konrad, 1998).

Internal Verifiers (IVs) are competent against occupational standards and are qualified as an Internal Verifier i.e. they hold D34 (Internally verify the assessment process). IVs ensure the quality of assessment judgements and processes within an approved centre. They select and train assessors and monitor assessment by sampling candidate evidence of competence and maintaining and developing the quality of assessment and verification documentation. They authorise requests for the award of NVQs and ensure equality of access to assessment

for all candidates (Konrad, 1998).

External Verifiers (EVs) are appointed by Awarding Bodies. EVs are required to hold the D35 unit (Externally verify the assessment process). EVs ensure the reliability and validity of the assessment and internal verification process, across NVQ centres. EVs also provide guidance and support to centres. The process of external verification involves sampling assessment and verification practice, providing feedback and an EV report to the centre and the appropriate Awarding Body. The verification chain, specifications and evidence of achievement, are viewed by some, for example, Eraut et al. (1996) to be paper dominated. Eraut et al. (1996) conclude that ultimately the cornerstones of the NVQ, i.e. specification and verification, do not necessarily guarantee standards and fairness. The *Centre Co-Ordinator* is the person who remains in contact with the Awarding Body.

The main way of securing the reliability of assessment of NVQs was considered to be writing detailed standards. Wolf (1998) argued that the belief in standards being secure because they are written down is a fallacy. She points out that there is a body of knowledge emphasising the role of tacit understanding, professional judgement and assessor networks in establishing standards (Wolf, 1998). This will be considered in more detail later.

There has been some research into the consistency of NVQ assessors' judgements i.e. the reliability of NVQs, see for example, Eraut et al. (1996) and Murphy et al. (1995). The reliability of NVQs depends upon the consistency of assessor judgement. These studies have tended to be quite negative about the reliability of NVQs. This raises the question of how the consistency of judgement might be improved. Eraut and Steadman (1998) and Konrad (1998) have separately suggested that reliability might be improved if there was more of a tighter network between assessors or a community of practice. One area which has not been researched and where it might be possible to develop tight networks is within NVQ centres. NVQ centres are required to undertake activities to standardise assessor judgement but anecdotal evidence suggests that this is not the case. With this in mind a survey of NVQ centres was undertaken to investigate whether centres engaged in standardisation activities, and if so what activities did they undertake and what were the advantages and disadvantages of these activities? The factors preventing centres from undertaking standardisation activities and some other issues were also investigated. It was hoped that the information might be used to provide guidance about standardisation to centres. This paper will focus upon the standardisation activities undertaken by centres, their advantages and disadvantages, and whether the activities tally with those suggested in the literature.

To give a focus for the study Retail Operations Level 2 Scheme Code 426, unit 2 (Meeting Customers' needs for information and advice) offered by Oxford Cambridge and RSA Awarding Body (OCR) was used as a case study. This scheme was chosen in consultation with OCR staff. It was thought that this vocational area and unit might be a useful way to

generalise, although it is acknowledged that any cross-vocational generalisations are limited.

There is already significant literature about standardising assessor judgement in areas other than NVQs. In the area of language testing there is evidence that training can bring examiners' differences in severity to a tolerable level but that it cannot eliminate differences in severity. It can also make examiners more consistent in their individual marking (Weigle, 1998; Stahl and Lunz, 1991; Lunz et al., 1990). Ruth and Murphy (1988) and later Weigle (1999) provide evidence that inexperienced language testing examiners were more severe than experienced examiners. Ham (2001) studied moderation systems in New Zealand where the education system is outcomes based, much like NVQs. He found that moderator and assessor experience was more important than subject experience for consistency of judgement within centres. Wigglesworth (1993) experimented with the feedback provided to language testing examiners as part of the training and standardisation process. She found some evidence that biases were reduced following feedback and that examiner-consistency improved. Barrett (2000) found that for a university level Communication and Media examination there were unacceptable levels of inter-rater reliability (consistency between assessors) and there was also one examiner who was too lenient. They also looked at other types of error - the halo effect, the central tendency effect and restriction of range. They found that one particular examiner was particularly free of error. They suggested that this was a matter of ownership and that increased ownership might increase inter-rater and intra-rater reliability (consistency within each assessor). Black et al. (1989) reported that in Scottish National Certificate modules there was a Communication module. The assessors had found it difficult to interpret the assessment criteria so they founded a network where standards were discussed. This led to a common understanding of the criteria which in turn led to an improvement in reliability. On the basis of this study and assessment literature like Wood (1991) and Dunbar et al. (1991), Wolf (1995) argues that discussion about candidates' work and assessment criteria between tight networks of assessors facilitates the reliability of assessor judgements, hence the title of the paper.

These principles of ownership, feedback, discussion and a tight network can be summed up by the principle that two heads are better than one. They also fit well with Konrad's suggestion that NVQ reliability might be improved by a community of practice constituting Internal Verifiers (IVs). In the literature about communities of practice it is explained that they facilitate learning with the result that more experienced members of the group pass the social practice on to newcomers who in turn might contribute new ideas to the social practice of the group. In this way members of all levels of experience have some ownership of the social practice and the social practice itself is dynamic (Wenger, 1998). The community of practice literature fits well with literature about standards as Cresswell (2000) explains when he states that standards are socially constructed and that applying standards is a form of social practice.

If NVQ assessors are to improve assessment through their standardisation activities then it would be beneficial if the activities included the factors which the literature suggests improve the consistency of judgement.

Method

The questionnaire was addressed to the contact listed on OCR records. In many cases the contact role was *The Centre Co-Ordinator* who was the intended recipient. The questionnaire was circulated in early to mid June and a reminder questionnaire was sent on to centres from whom a questionnaire had not yet been received.

All centres who were registered with OCR to offer Retail Operations NVQ Level 2 Scheme Code and who were located in Ireland, England, Scotland and Wales were sent a questionnaire. In total 570 centres were mailed a questionnaire and136 completed questionnaires were returned. Given that not all centres registered with OCR are necessarily active centres and that some centres had idiosyncrasies (see caveats) that prevented an accurate estimation of the response rate, there was a response rate of approximately 24% -mirroring a similar response rate to Eraut et al. (1996). This estimate of the response rate excludes the centres who replied stating that their centre was currently inactive.

Analysis

Analysing the responses to the questionnaire involved both quantitative and qualitative approaches.

In the quantitative analysis the frequency with which each response is given e.g. the number of centres which responded 'yes' and the number who responded 'no' is reported. There were some questions where it was possible for a centre to give more than one response e.g. '5iv Who attended the meetings (tick all boxes that apply)'. For these questions the number of centres which gave each response along with the frequency of the patterns of responses e.g. the number of centres which ticked both 'full time assessor(s)' and 'part time assessor(s)' is reported.

The text of the responses to the open ended questions were divided into meaning units (phrases and / or sentences of text which are meaningful on their own) and analysed question by question. The data and in some cases the relevant research literature (reviewed above) were used to identify appropriate categories. Each category was given a code. Therefore categories and codes were developed both inductively and deductively. There was some double coding and some responses constituted more than one meaning unit. Hence there is no one to one relationship between the frequency of coded meaning units and the number of centres which responded. The categories for the different questions seem

to have some similarities and overlaps. This is not surprising given that the information is all about a similar topic.

Results and Discussion

The collective response to each question is treated separately in the section below.
The following frequencies and statistics should be treated with some caution. For example, there is one chain of centres who have a policy to answer any questionnaires only through head office as all procedures are standardised. This information emerged as one centre rang to explain why they would not be returning a questionnaire. In the case of this chain it will be difficult to investigate anything other than centralised policy, the details of practice in individual centres will not be revealed. So there was only one response for that particular chain of centres. Given such idiosyncrasies in the organisation of some centres the frequencies cannot be read as directly representative of the Retail Centres.

As the focus of this paper is the standardisation activities undertaken by centres, their advantages and disadvantages, only the responses to questions which focused upon these issues are reported.

Table I. Responses to question 1i -
In the past two years has your centre undertaken activities to standardise assessor judgement?

Response	Frequency	Percent
Yes	110	82.1
No	24	17.9

The majority of centres had undertaken activities to standardise assessor judgements. This was a reassuring finding considering that there was anecdotal evidence that centres do not standardise assessor judgement. On the other hand it could be that out of the 570 centres that received a questionnaire the centres that tend to standardise returned the questionnaire.

Table III Responses to question 5i -
Do you use team meetings to standardise assessor judgements?

Response	Frequency	Percent
Yes	103	91.2
No	10	8.9

The majority of centres used team meetings to standardise assessor judgement. The 10 centres who did not use team meetings might use one to one contact between the Internal Verifier or Centre Co-ordinator and the assessors to standardise judgements. This approach was mentioned in some of the methods described in response to question 7. There were 113 centres answering question 5i as some centres did not follow the question routing.

The 103 respondents who answered 'yes' to question 5i were asked to complete questions 5ii, 5iii and 5iv. If the answer was 'no' then the respondents moved on to question 6i.

Table IV Responses to question 5ii -
How many meetings were held in the past two years?

Number of meetings	Frequency	Percent	Cumulative frequency	Cumulative percent
1- 5	17	20.0	17	20.0
6 - 10	23	27.1	40	47.1
11 - 15	13	15.3	53	62.4
16 - 20	7	8.2	60	70.6
21 - 25	20	23.5	80	94.1
26 - 30	1	1.2	81	95.3
31 - 35	0	0.0	81	95.3
36 - 40	2	2.4	83	97.6
41 - 45	1	1.2	84	98.8
46 or more	1	1.2	85	100.0

The number of team meetings held in the past two years varied considerably between centres. One centre held by far the most meetings. Table IV does not tell the whole story, 11 i.e. 10% of the centres had bi-monthly and 17 (16.5%) of the centres had monthly meetings. This fits with some of the responses to later questions which referred to monthly meetings: *Ongoing training is carried out continuously via monthly assessor meetings/IV meetings/assessor workshops.* The variety in the number of meetings might be due to the different length of meetings in different centres and the unique situation of each centre. There was missing data for question 5ii for 3 centres.

Responses to question 5iii -
Were the meeting(s) well attended?

All but one of the 102 respondents gave a positive answer to this question. This high number might be partly due to the positive answer being the socially desirable answer. We return to the issue of attendance at standardisation activities later in the report – see question 7. There was missing data from 1 centre for question 5iii.

Table V Responses to question 5iv -
Who attended the meeting (tick all boxes that apply)?

Response	Frequency	Percent
Assessor(s) working as line managers (1)	55	14.9
Full-time assessor(s) (2)	97	26.3
Part-time assessor(s) (3)	58	15.7
Assessor(s) from dispersed locations within the centre (4)	4	11.6
Assessor(s) no fixed location within the centre (5)	7	1.9
Experienced assessor(s) (6)	75	20.3
Inexperienced assessor(s) (7)	34	9.2

Note: The number in brackets after the response e.g. Full time assessor(s) (2) is the code that the response was given and is used in Table VI below.

The most frequently occurring responses were full-time and experienced assessors. There are also a good proportion of part-time assessors and assessors who are line–managers. It seems that there was a broad cross section of assessors who are involved in standardisation. This is positive given that a sense of ownership in standards and their application can lead to an improvement in the reliability of assessment (Barrett, 2000).

Table V cannot be used to identify how many people there were in total at a meeting or how well attended the meetings were. This is because we do not know how many assessors there were from each centre and which categories each assessor fits into. Nor can the figures be used to identify the proportion of assessors in each category who attended each meeting. This question was used to give an indication of the cross section of staff that were involved in standardisation. Table VI below gives a better feel for the types of assessors who attend the same meeting.

Table VI The mixture of types of assessors who attended the meetings

Responses	Frequency	Percent
1234	22	22.0
1236	11	11.0
2367	9	9.0
26	9	9.0
1246	8	8.0
267	7	7.0
236	6	6.0
1267	5	5.0
2467	5	5.0
126	3	3.0
367	2	2.0
1235	1	1.0
1245	1	1.0
1256	1	1.0
1345	1	1.0
167	1	1.0
2345	1	1.0
2346	1	1.0
235	1	1.0
2356	1	1.0
27	1	1.0
346	1	1.0
36	1	1.0
67	1	1.0

By far the most frequently occurring group of assessors to attend a meeting were made up of assessor(s) working as line-managers, full-time and part-time assessors and assessors from dispersed location(s) within the centre. There were half as many occurrences of a group of assessor(s) working as line managers, full-time assessors, part-time assessors and experienced assessors attending meetings.

The qualitative responses to question 7 concerned the method(s) used to standardise assessor judgement. There were a number of standardisation activities described in the questionnaire responses that tally with the methods of improving reliability found in the literature. The most frequently mentioned of these methods was ‘Feedback given by the IV to the assessor(s) on their assessment judgements’ occurring 33 times. Although giving feedback has been identified as useful it might be that the hierarchical relationship between an IV and an assessor might discourage learning how to apply the assessment criteria, given that Wenger (1998) has argued that flat hierarchies facilitate learning. ‘Feedback given by

the assessor(s) to other assessor(s) on their assessment judgements' occurred 15 times. The later might be more conducive to learning than the former. It was mentioned 24 times that 'assessors share examples of evidence or candidates' portfolios'. Additionally there were 20 comments about various forms of 'training' being given to assessors, 18 responses about 'discussion between the IV and assessor(s) about their assessment judgements and 19 responses about 'discussion between assessor(s) about their assessment judgements'. All these activities fit well with the literature and were fairly popular. It is positive that the principles found in the literature are practised in some NVQ centres. However there was also the category of 'a tight network of assessors or communication between assessors' that only occurred 7 times. This category fits well with the notion of a community of practice facilitating reliability of assessment judgements and a tight network is something that Wolf (1995) argues is essential for reliability. This also suggests that it would be beneficial if more NVQ centres focused upon developing a tight network of assessors than is currently the case. It would also be beneficial if there were more centres engaged in the discussion of assessment judgements of examples of candidates' work and also offered their assessors training as well as gave providing feedback about assessment judgements to assessors.

In addition to the activities found in the literature the centres also used other standardisation activities. To date there does not appear to be any evidence to illustrate whether these activities are beneficial for improving reliability. This might be a worthwhile area of future research. The other activities were the 'discussion of examples of assessment methods, how assessment objectives and different types of evidence might be used' mentioned 39 times, and centres 'agreed best and / or bad practice through discussion of issues'. It seems that these sessions provide a time of general discussion and reflection that is undoubtedly a useful activity.

Another method was 'regular meetings'. The issue of team meetings has already been discussed in relation to questions 5i to 5iv. There were 13 mentions of both 'The IV checking decisions' and 'Involving the EV in various ways e.g. discussion with the EV and / or using the EV report', 8 occurrences of 'using assessment forms or standardising paperwork and / or evidence' and 4 responses about 'using quality assurance procedures'. Clearly, NVQ centres use a greater variety of methods to help standardise assessors' judgements than the literature recommends. Indeed these activities are unlikely to have been found in any literature about other forms of assessment as some are specific to the NVQ situation, illustrating that the NVQ centres are using the resources that are available to them.

The second part of question 7 asked respondents to consider the advantages of the method used. Of the advantages which were contained in the responses 'communication between assessors and general discussion between assessors' was mentioned 22 times. This can be related to the research literature suggesting that communication within a tight network of assessors can facilitate the reliability of judgements. However if the discussion is too general and not specifically about assessment decisions then it might be less useful. Some

of the advantages which were derived from the literature but which occurred less frequently were:-

1) 'assessors are involved which might give a sense of ownership' - occurred 9 times;
2) 'assessors receive good feedback' - occurred 8 times;
3) 'assessors reach agreement or are encouraged to reach agreement on assessment judgements' - occurred 8 times;
4) 'assessment judgements are discussed' - occurred 7 times;
5) 'inexperienced assessors are involved in the standardisation' - occurred 6 times;
6) 'experienced assessors are involved in the standardisation' - occurred 3 times.

Number 3) is perhaps the most important advantage but it is not explicitly mentioned by a good many centres, this might be because it is too obvious an advantage! Arguably it is a little disappointing that some of the factors mentioned in the literature are considered to be advantages by so few centres. However this could be because the methods which work in other contexts like language testing might not be practical or advantageous in the NVQ situation. Evaluating methods of standardisation in centres is a possible area of future research. The centres did not mention having trained assessors as an advantage of the methods used but this was the only factor in the literature expected to improve reliability which the centres did not refer too. Also centres must see having trained assessors as an advantage otherwise they would not train them. In summary, although the factors in the literature were not mentioned by many centres most of the factors are acknowledged as advantages by NVQ providers.

One of the most popular advantages of the methods of standardisation that occurred 22 times was that 'assessors learn, e.g., broadens the assessors' experience'. Another equally popular advantage was 'supportive environment and good staff and candidate relations'. These and the other less frequently mentioned advantage are the points that are probably not unique to the NVQ sector, although they have not been covered by the literature about improving the consistency of assessor judgement.

The third part of question 7 asked about the disadvantages of the methods of standardisation. The only response that related to the literature on standardisation was that one disadvantage of some methods was 'a lack of discussion and / or agreement' mentioned 14 times. It is positive that centres recognise lack of discussion as a disadvantage given that it is a way of improving reliability. An area of more concern is that some disagreement between assessors in centres might be related to assessment judgements. Any disagreements about assessment judgements might be related to another disadvantage: 'assessment is subjective as it is based upon judgements and opinions' referred to 3 times. This goes back to Cresswell's (2000) discussion about standards that they are socially constructed and are a form of social practice. From this perspective, judgements about standards are always subjective.

The most often mentioned disadvantage given by the centres was that 'the time required to undertake standardisation / finding a time when everyone is available' was mentioned 47 times. This tallies with the responses to question 2 illustrating that a great deal of staff time is taken up in standardisation activities. This disadvantage was mentioned more often than any of the methods of standardisation listed in responses to question 7 and more times than any of the advantages of the standardisation activities. It is therefore of paramount importance that any methods of standardisation evaluated in centres in future research should be evaluated for their efficiency. Another disadvantage of some methods of standardisation is that 'standardisation is removed from assessment practice in various ways, so the validity of assessment judgements is compromised in standardisation exercises' mentioned 15 times. This is an important point, if assessors are inspecting candidates' portfolios in standardisation exercises then they do not have the same information at hand as the assessor and therefore are making different judgements. For example, the portfolio is really the only evidence that a candidate has achieved particular performance criteria. When assessors make assessments they do not just inspect a candidate's portfolio, they might also interview a candidate, see artefacts that the candidates have made and / or observe a candidate performing etc. Some standardisation activities might involve two assessors observing a candidate, or assessors observing one another interviewing a candidate. This might be a more valid method of standardisation, although it might prove intrusive. This point that there are sometimes too many assessors observing a candidate was mentioned by 2 centres. It is difficult to see how this disadvantage might be overcome.

One of the advantages with some methods was that there were positive staff relationships. Also, of course, there can be negative staff relationships. 14 centres made comments that were put in the category 'Relationships between staff might become negative e.g. if criticism is taken personally, not everyone wants to be involved in discussion and sharing'. It is unfortunate that discussion which is so useful for standardising judgements might also foster negative staff relationships. Obviously good staff relationships are needed for positive discussion and to build a community of practice. There were 12 centres claiming no disadvantages, as yet, found with their method of standardisation. This does suggest sound practice in the area of standardisation.

Seven centres mentioned that 'During the standardisation process assessors are trying to deal with too much diversity e.g. in the occupations represented at the meeting / the units to be standardised/candidates work'. In terms of there being a variety of different backgrounds Brown (1995) developed a language test for which assessors from different occupational groups gave candidates the same grade. This suggests that it is possible for assessors of different occupational backgrounds to reach an agreement. However Brown (1995) also argued that if the assessors from the different backgrounds had developed tests in their separate occupational groups then they would have developed different tests. There were 30 centres that reported both experienced and inexperienced assessors being involved in team meetings where standardisation was undertaken (see Table VI). It could be that some

centres tried to standardise across too many occupations and qualifications thus bringing too much diversity to standardisation activities. On the other hand there could be difficulties involved in engaging with a social practice like standardising assessment with people from a variety of backgrounds. If this is the case then a communities of practice approach would be to increase the amount of networking and discussion between the individuals involved. Of course this requires time which is in short supply and valuable. There were a small number of centres who mentioned that perhaps there was not sufficient diversity in the work they considered in standardisation. It might be useful to have a range of evidence and assessment methods with a group of assessors having the same qualification/unit for standardisation purposes.

In the responses to question 7, six centres mentioned that 'Assessors' level of experience can have various effects'. Wenger (1998) explained the positive contributions of experienced and inexperienced members to a community of practice. He also acknowledged that there might be some difficulties when new members joined as it could upset the status quo. The comments suggested that mentoring new assessors can take time and they may feel threatened if they cannot identify poor/good practice. The evidence from this questionnaire suggests that the impact of inexperienced assessors does not appear to affect the status quo. The idea of mentoring inexperienced assessors as suggested by one centre would be a good way of offering newcomers a way into the community of practice in a centre.

Discussion and Conclusions

Of the centres that responded to the questionnaire a good number undertake standardisation; 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 other contexts. This suggests that the methods of standardisation and ways of facilitating reliability which have been found in other contexts are practised in the context of centres who offer Retail Operations. What remains is to test which of these methods have positive effects in the NVQ context. This is likely to be the subject for further research when methods of standardisation are tested in centres and evaluated. The methods listed above need to be developed before they are tested in centres. For example, standardisation activities cannot be based simply upon 'Feedback given by the IV to the assessor(s) on the assessment judgements'. Before these methods can be tested in centres EVs and / or IVs need to develop one or more standardisation activities based upon the focus group and questionnaire responses.

Whilst the questionnaire responses give a good overview of the activities that are undertaken to standardise assessment judgements they do not explain which activities take place together and which work well in combination. For example, it could be that 'feedback given by assessor(s) to other assessor(s) on their assessment judgements' is often combined with 'training'. It might also be the case that as a combination they are powerful in

improving reliability. It would take further analysis of the questionnaire data to address the former issue and further research to address the latter.

Unfortunately this analysis does not:-

- link the methods of standardisation to particular advantages and disadvantages;
- indicate how methods of standardisation might vary with assessment method.

Further research might include undertaking some interviews at one or two centres as case studies to verify the contents of the questionnaires. A principal components analysis of the questionnaire data might indicate what activities were undertaken together in centres. Additionally it would be useful to test whether the methods of standardisation tested in other contexts are facilitating the reliability of assessment in the NVQ context.

Acknowledgements

The author is grateful to administrative staff at UCLES for involvement in the administration of the questionnaire, research colleagues and NVQ experts for comments on the questionnaire and Mike Lewis for his involvement in the data handling.

Bibliography

BARRETT, S. (2000) *HECS LOTTO: Does Marker Variability make examinations a lottery?* Division of Business and Enterprise, University of South Australia. www.aare.edu.au/99pap/bar99789.htm

BLACK, J. H., HALL, J., MARTIN, S. & YATES, J. (1989) *The Quality of Assessments: Case Studies in the National Certificate*. (Edinburgh, Scottish Council for Research in Education).

BROWN, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test, *Language Testing*, (12) 1 pp. 1-15.

CRESSWELL, M. J. (2000) The Role of Public Examinations in Defining and Monitoring Standards, in H. GOLDSTEIN AND A. HEATH (2000) *Educational Standards* (New York, Oxford University Press).

DUNBAR, S. B. KORETZ, D. M. AND HOOVER, H.D. (1991) Quality control in the development and use of performance assessments, *Applied Measurement in Education*, (4) pp.289-304.

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

ERAUT, M., STEADMAN, S., TRILL, J. & PARKES, J. (1996) *The Assessment of NVQs*. *Research Report Number 4* (Brighton, University of Sussex).

HAM, V. (2001) *Maintaining National Standards in Standards Based Assessment: The New Zealand Experience*, A paper presented at the British Educational Research Association Conference, University of Leeds, 13th – 15th September 2001.

KONRAD, J. (1998) *Assessment and Verification of National Vocational Qualifications: a European quality perspective*. Education-line www.leeds.ac.uk/educol/index.html.
LUNZ, M.E., WRIGHT B.D. AND LINACRE, J.M. (1990) Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* , (3) pp.331-45.
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 (Nottingham, School of Education, University of Nottingham.
RUTH, L. & MURPHY, S., (1988) Designing writing tasks for the assessment of writing (Norwood, NJ: Ablex Publishing Corp).
STAHL, J. A. AND LUNZ, M.E. (1991) *Judge performance reports: media and message*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
WEIGLE, S. (1998) Using FACETS to model rater training effects, *Language Testing*, (15) 2, pp. 263-287.
WEIGLE, S. (1999) Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches, *Assessing Writing*, (6) 2, pp.145-178.
WENGER, E. (1998) *Communities of Practice Learning, meaning and identity* (Cambridge, Cambridge University Press).
WIGGLESWORTH, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing,* (10) (3), pp.305-35.
WOLF, A. (1995) *Competence Based Assessment*. (Milton Keynes, Open University Press).
WOLF, A. (1998) Portfolio assessment as national policy: the National Council for Vocational Qualifications and its quest for a pedagogical revolution. *Assessment in Education, Principles, Policy and Practice*. (5) 3, pp.413-445.
WOOD, R. (1991) *Assessment and Testing: a survey of research* (Cambridge: Cambridge University Press).