CAMBRIDGE ASSESSMENT

Towards a methodology for evaluating the equivalency of demands in vocational assessments between colleges/training providers

Victoria Crisp and Nadežda Novaković

Research Division, Cambridge Assessment

A paper presented at the International Association for Educational Assessment Annual Conference, September 2008, Cambridge, UK.

Contact:
Victoria Crisp
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK
Direct dial. +44 (0)1223 553805
Email: crisp.v@cambridgeassessment.org.uk

Towards a methodology for evaluating the equivalency of demands in vocational assessments between colleges/training providers

Victoria Crisp and Nadežda Novaković
Cambridge Assessment

Abstract

In many countries and contexts, vocational skills tend to be assessed via observations (or other evidence) of the practical performance of appropriate tasks. In the case of national qualifications, although the nature of the tasks may be broadly set out within the qualification's assessment guidance, colleges or training providers often determine the particular tasks used. This system allows for the assessments to be as authentic as possible. However, it is possible that aspects of the assessment contexts (e.g. task instructions and guidance, level of support provided by the tutor) could lead to differences in task demands between colleges/training providers. Unless any such differences are accounted for in assessment judgements, they could pose a threat to assessment reliability, and hence, to validity. This paper will describe the trialling of a possible methodology for comparing the demands of vocational assessments between colleges/training providers. Judges with relevant experience revised a framework of cognitive demands and then made paired comparisons of assessment materials from a vocationally-related administration qualification. Additionally, some tutors and students were interviewed to gain insight into the assessment contexts. The methodology was generally successful but difficulties and possible improvements will be discussed.

Introduction

The National Qualifications Framework (QCA 2004) links general, vocational and vocationally-related qualifications and is designed to unify the UK qualifications system by indicating equivalency of different qualifications. Thus, the consistency of assessment demands and assessment judgements is important since low levels of assessment consistency could pose a threat to the mutual recognition of different qualifications. It is therefore necessary to have methodologies available to investigate such issues. While some mature methodologies exist and have been used to compare general qualifications (see Newton et al. 2007), which mostly involve written examinations, investigating such issues in relation to vocational skills is more problematic, and limited attempts have been made in this area. This paper proposes a methodology for comparing the demands of vocational assessments which would provide measures that can be statistically analysed.

Vocational skills are often evaluated by practical assessments rather than written examinations because students can be assessed whilst conducting tasks that would be undertaken in the relevant job role. However, this leads to the dispersed assessment of students across a large number of workplaces, training providers and/or colleges. Awarding Bodies aim to encourage equivalency of assessments across colleges nationally by providing guidance and support materials and opportunities for training. For many nationally recognised vocational qualifications, moderation or verification procedures are also in place to standardise assessments. However, the dispersed nature of such assessments combined with the cognitive load of contextualization (i.e. that generic criteria must be interpreted within different

assessment contexts, see Oates 2004) leaves the system open to greater potential risks of lower consistency of assessment tasks, assessment demands and assessment judgements.

Although assessment demands represent only one aspect of examination standards, having a method that would allow us to compare the demands of different vocational assessments could contribute to the reliability and robustness of assessments by giving us a way to monitor this aspect of standards not only from year to year but across different contexts and different qualifications. This paper describes a methodology for comparing the demands of vocational assessments between colleges or training providers.

To our knowledge, there have been no attempts to date to devise a suitable method for comparing assessment demands, specifically within the context of vocational qualifications. However, there has been some research into the reliability of vocational assessments. The relative paucity of research in this domain could be because, as Murphy et al. (1995) assert, it is more difficult to measure the reliability of performance-type assessments than of examinations as the traditional statistical procedures for measuring reliability are not appropriate for assessing work-based learning (see also Benett, 1993).

Most of the studies conducted so far have shown that researching the reliability of vocational qualifications is beset by many issues not present in the context of general examinations. Assessors in vocational contexts very often need to judge candidate skills and competence based on the evidence of how they perform on specific tasks in specific settings, although the research literature suggests that there may not be a direct transfer of skills from one context or situation to another (Boreham et al. 2002, Wolf and Silver 1990, Oates 1992). This is closely related to the fact that candidates' skills may be assessed either through simulated or authentic tasks or activities, which may affect the validity of the assessment process (Tolley et al. 2003, Murphy et al. 1995, Wolf 1995).

Furthermore, the opacity of vocational standards (Beaumont 1995) leaves them open for individual interpretation, which consequently leads to situations where the assessment of candidates' competence is influenced by individual assessors' subjective perceptions of the assessment requirements or their own ideas of what constitutes a competent candidate. In the complex context of the validity, reliability and comparability of the assessment of vocational skills, it was necessary to narrow the focus of our investigation to just one strand of these issues: *'how can we evaluate the consistency of task demands in the assessment of vocational skills?'* We were aware that in focusing in this way, the proposed methodology would be able to provide only one part of the picture regarding the comparability of standards, given that tutors' assessment judgements and moderation procedures can compensate for any differences in demands[1].

[1] In this development and piloting our focus was on the underpinning cognitive demands that assessment tasks place on students and not on how difficult a task may turn out to be for a particular group of students (although the two concepts are, of course, interlinked). We used a definition of demands proposed by Pollitt, Ahmed and Crisp (2007) which defines demands as "separable, but not wholly discrete, skills or skill sets that are presumed to determine the relative difficulty of examination tasks and are intentionally included in examinations" (p.196).

Methodology development

Pollitt, Ahmed and Crisp (2007) recently reviewed the methodologies used in previous research to compare or measure the demands of assessments, including in Awarding Body comparability studies in the UK. Such studies compare the demands and standards of (almost exclusively) general qualifications between examination boards.

Over time, a range of different methods have been used to evaluate assessment demands in comparability studies. For example, Massey (1979) attempted to evaluate the 'inherent difficulty' or 'complexity of demand' of English syllabuses by asking judges whether examination papers were 'relatively demanding', 'relatively undemanding' or 'average' in relation to a number of aspects of reading and writing demand. Somewhat similarly, Houston (1981) asked examiners to categorise different types of demands in Economics examinations as 'excessive', 'appropriate' or 'insufficient'. Later, McLone and Patrick (1990) investigated factors affecting demands by asking judges to rate the demands of mathematics examination questions on a scale from 0 to 3 on a number of questions and to rank-order syllabuses according to perceived demand.

A number of comparability studies in the early to mid 1990s (e.g. Jones 1993, Stobart et al. 1994, Ratcliffe 1994, Gray 1995, Phillips and Adams 1995) also asked judges to rate aspects of demands but this time on a scale of 1 (easy) to 5 (demanding). The dimensions of demands rated were usually drawn from research by Pollitt et al. (1985). The general pattern for the factors used was often:

- 'content' (in relation to the requirements and demands of the syllabus and the associated question papers) or 'subject/concept difficulty'
- 'skills and processes'
- 'structure and manageability of the question papers' (question difficulty, language, layout, context, etc.) or 'question difficulty'
- 'practical skills' (particularly in relation to fieldwork) or 'using and applying' (in relation to coursework)

The ratings were used to compare the demands of syllabuses. This often resulted in identifying that certain specifications were slightly more or less demanding in certain respects.

More recent comparability studies (e.g. Arlett 2002, Edwards and Adams 2002, Edwards and Adams 2003, Greatorex, Elliott and Bell 2002, Guthrie 2003) have tended to use a slightly different methodology to compare demands. This involves an initial phase drawing on Kelly's Repertory Grid technique for construct elicitation (Kelly 1955). In this stage a number of relevant experts are asked to compare examination materials from pairs of specifications and write down similarities and differences in the demands placed on candidates. These ideas are pooled and discussed and an agreed list of features of the assessments that affect demands is compiled. Gray (2000) describes this as enabling examiners to form their own ideas of what constitutes demand in order to derive constructs and define a scale of demands. A larger group of expert judges is then asked to rate the materials from each specification under scrutiny in terms of how demanding it is in relation to each feature/construct on the list via a questionnaire. Ratings are usually from 1 to 5 or 1 to 7. Mean ratings can then be compared between specifications. The judges were generally reported to consider the methodology satisfactory (e.g. Phillips and Adams, 1995). However, a number of methodological difficulties have been raised. Firstly, the

procedure can result in a large number of features to be rated. Linked to this, the features to be rated have varied in the extent to which they are explicitly about demand, to what degree it is known how the factors really affect demands and whether they affect all students in the same way. Secondly, there are a number of problems associated with the use of numerical scales. Phillips and Adams (1995) report that some raters found the scales to be limiting and that judges wished to expand the scale with '+ ' and '-' subdivisions. There have also been reports of difficulties with interpreting the questions for those not involved in the derivation of constructs and with interpreting the descriptors that mark the ends of scales (Fearnley 2000). Additionally, there is a risk that judges who give the same rating will not necessarily 'mean the same thing' by that particular rating, perhaps basing the centre point of the scale on the level of demand of the specification with which they are most familiar. This may result in the average of judges' ratings not being very meaningful. Some judges have reported that they are unsure of whether they are making consistent rating judgements. Additionally, it is difficult to know the size of differences between ratings in real terms.

Whilst methods using construct elicitation and rating scales are generally considered effective (e.g. Fearnley 2000, Arlett 2002), an improvement on earlier methods (Gray 2000) and have provided useful insights as long as interpretations are made with some caution (e.g. Adams and Pinot de Moira, 2000), such methods do have methodological weaknesses. Consequently, a methodology drawing on this and two other methods sometimes used to compare qualifications was designed.

The first additional method, a paired comparisons method (originally proposed by Thurstone 1927), has previously been used to investigate the equivalency of student work achieving particular grades across different awarding bodies as a part of comparability studies. In this method, judges compare the work of two students at a time and decide which is 'better'. The outcomes of numerous comparisons can be analysed using Rasch analysis and allow the relative quality of student work to be ascertained, and hence any difference in the standard to be identified. Using relative judgements avoids the problems associated with rating scales. The use of paired judgements can also allow year on year comparability to be investigated.

The proposed methodology also draws on a set of dimensions of cognitive demands previously defined for use in comparing the demands of assessment materials. Drawing on work by Edwards and Dall'Alba (1981), and by working with examiners, Hughes, Pollitt and Ahmed (1998) developed a scale of cognitive demands that could be used to rate the demands which examinations place on candidates in terms of each of a number of dimensions (or types) of demand on a scale of 1 to 5, and thus compare questions. The demand dimensions were:

- Complexity – the number of components or operations or ideas involved in a task and the links between them
- Resources – the use of data and information (including the student's own internal resources)
- Abstractness – the extent to which the student must deal with ideas rather than concrete objects
- Task strategy – the extent to which the student devises (or selects) and maintains a strategy for tackling the question
- Response strategy – the extent to which students have to organise their own response

This has been used in recent studies to compare the demands of assessments between subjects and between different levels of qualification (e.g. between GCSE and A level) in the same subject (see, for example, QCA 2008). However, this method uses a rating scale for each dimension of demands and thus has some of the previously mentioned methodological challenges in relation to ratings. Consequently, the proposed methodology begins with the dimensions of the scale of cognitive demands defined by Hughes, Pollitt and Ahmed, revising these as necessary using a Kelly-type construct elicitation method and then asks judges to make paired comparisons of assessment materials in terms of each dimension or type of demand. As the demand dimensions in Hughes, Pollitt and Ahmed's scale have not previously been used with vocationally-orientated practical assessments, the proposed methodology includes a stage of tailoring or revising this framework.

Proposed methodology

The proposed methodology brings together aspects of the methodologies described above, but attempts to avoid some of the difficulties associated with them. The methodology is outlined below.

Preparation and construct elicitation

- Appropriate judges are provided with materials illustrating the assessment tasks undertaken by students from a number of colleges/training providers running the course and are asked to familiarise themselves with these materials. The judges are also provided with a description of the definition of demands that will be used in the study.
- Individually, judges are asked to compare two pairs of colleges' materials in turn (e.g. compare those from College A with those from College B, etc.) and to identify and write down similarities and differences between them in terms of the demands that they place on students.
- The Hughes, Pollitt and Ahmed (1998) scale of cognitive demands is provided to judges but with each dimension presented as a continuum from 'less demanding' to 'more demanding' rather than as rating scales. Judges are provided with a blank version of this framework of demands (based on the Hughes, Pollitt and Ahmed dimensions) with space to add examples of how the dimensions relate to the specific qualification and space to add additional or alternative dimensions.
- Judges are then asked to consider whether and where the types of demands they have identified, through focussing on similarities and differences, slot into the framework of demands. They are asked to write in specific examples and to suggest additional dimensions where necessary.

Revision of the demands framework

- The completed frameworks are then compiled into a revised framework either by the researchers or by a small number of appropriate experts.

Paired comparisons of materials

- Judges attend a meeting where the revised demands framework is discussed with the aim of ensuring a shared understanding of the dimensions.
- Judges work individually and compare pairs of colleges' assessment materials in turn and judge for each pair which is more demanding in terms of each of the demand dimensions and then overall.

Interviews at colleges

- Interviewing tutors and students involved in the course at a number of colleges (preferably the same colleges as those whose assessment task materials are used in the stages above) would provide additional insight into the context of the assessments. Questions could be targeted to obtain information regarding the kinds of tasks conducted, which elements, if any, were conducted during work experience, the degree of detail and structure of tasks, the extent of support provided by tutors or workplace supervisors, and how evidence for assessment is gathered.

Framework for analysis

The paired comparison data can be analysed in two ways. If one has a complete or near complete design, calculating the frequencies with which the assessment materials from one college are judged as 'more demanding' than another college's assessments will provide a good indication of the relative demands of the different sets of materials.

Otherwise, the data can be analysed using a multi-faceted Rasch (MFR) model (Linacre 1989) and the FACETS software (Linacre 2007, version 3.60.0). Rasch analysis is a form of item response theory, which can provide more robust estimated measures of demand based on the paired comparison data. It takes into account the various sources of variability (for example, colleges, judges and types of demands) when calculating the size of the differences in demand between different sets of materials. The analysis also provides *separation index* statistics, which suggest the extent to which one can be reasonably sure that the measures obtained in the analysis are 'separated' from one another (i.e. different to each other).

Piloting the proposed methodology

Assessment context and method

The methodology described above has been piloted with two assessment units that form part of a vocationally-related qualification in administration (Crisp and Novakovic in press). This qualification tends to be taken by students at Sixth Form college (age 16-18 years) and by adult learners attending Further Education colleges. The course is usually college-based although many students undertake some work experience as part of the course. The two units used in the piloting are assessed by college tutors, and sometimes partly by workplace supervisors. One of the units assesses working with customers and colleagues and requires candidates to undertake group tasks (e.g. planning a charity event) to assess teamwork, and individual demonstrations of customer service skills, such as taking telephone calls and dealing with face-to-face enquiries (sometimes assessed via simulated activities). The other unit assesses the use of office procedures (e.g. photocopying, mail handling, stock control). For both units, students complete tasks over a period of time and then write review reports on how they went about tasks and their own strengths and weaknesses in conducting them. These reports, along with witness statements completed by college tutors or workplace supervisors, are submitted for external checking (moderation) by Awarding Body appointed examiners (moderators). Tutors decide when a student has met all necessary criteria and only submit folders of evidence when a student is judged to have done so.

Fifteen judges (tutors, examiners and moderators) were involved in the study, which was conducted as described in the 'Proposed methodology' section above. Judgements were made regarding the assessment tasks used in five colleges. The materials used to provide the judges with information on the assessments included written task outlines where these could be obtained (this was possible for three of the five colleges) and examples of student submission folders for each of the two units for each college. The assessments of both units within a college were treated together.

In this study, few additional types of demands were suggested and so just the dimensions of the original Hughes, Pollitt and Ahmed (1998) framework were used. However, the addition of examples of the ways that the different types of demands differ in the administration assessments was helpful in making the framework more specific for the judges.

Interviews were undertaken with the tutor and with two or three students in each of four colleges teaching the course. Unfortunately, it was not possible to visit the same colleges whose materials were being used in the paired comparisons.

Findings from Rasch analyses

In this illustration we will present only the Rasch results for reasons of brevity. Using frequencies of judgements produced a very similar pattern of results in most cases.

Rasch analyses were conducted firstly by including the judgements in relation to each type of demand as a separate 'facet' within one overall Rasch model to provide overall measures of demand, and secondly, a separate Rasch analysis was conducted for each type of demand and for judgements of overall demand.

Figure 1 represents estimates of the demand of different colleges' materials measured in logits (log odds units), as obtained from the analysis which treated types of demand as different facets contributing to an overall model. The measures show which college's materials were more demanding (positive values) and which were less demanding (negative values). The most demanding assessment materials were from College A and the least demanding from College E. The separation index was 5.20 (reliability of 0.96), indicating that less than 4% of the variance between the observed demands is due to measurement error; i.e. these differences are meaningful. However, the differences in overall demands appear to be fairly small.

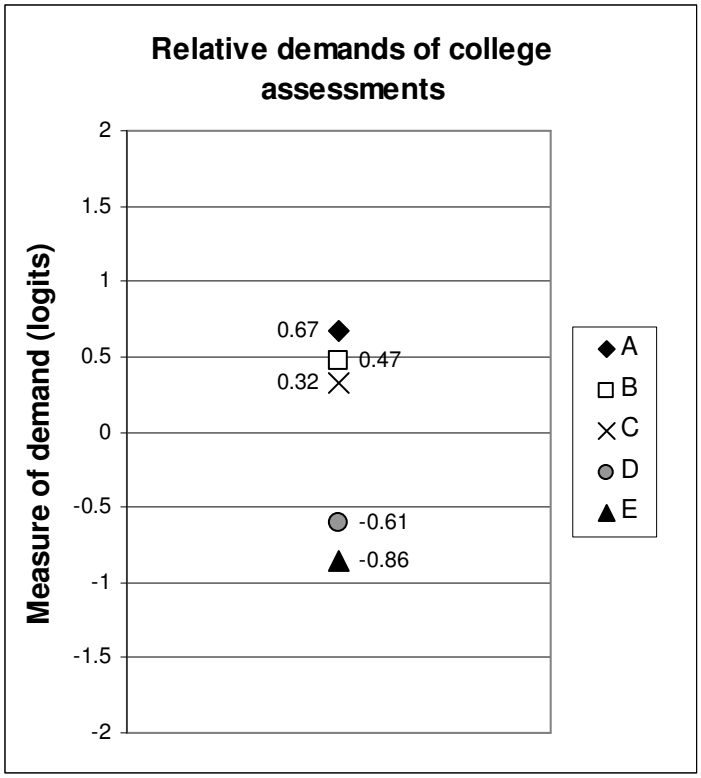

Figure 1. Estimates of relative demand overall based on multi-faceted Rasch analysis including each type of demand.

Figure 2 shows the results of Rasch analyses run on each of the judgment criteria separately, as well as on the judges' decisions about the overall demand of college materials.

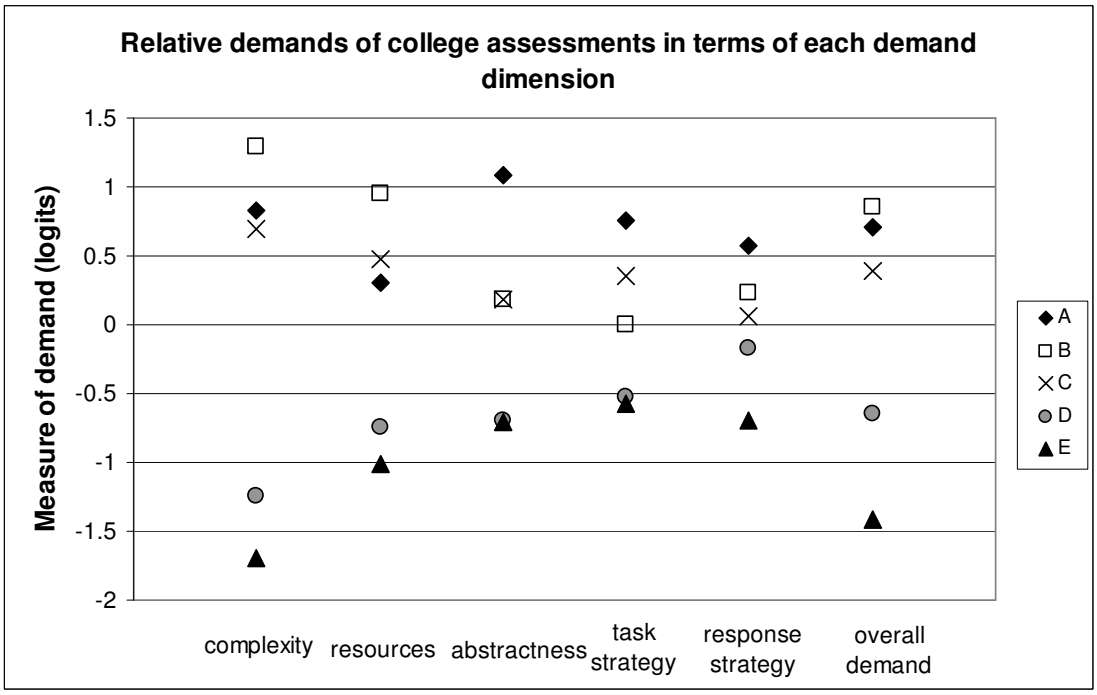

Figure 2. Estimates of relative demand for each type of demand based on separate Rasch analyses for each one

The assessment materials from Colleges D and E were generally judged to be less demanding than those from the other three colleges on each of the demand dimensions and overall but with variations in the rank-order of demand for some dimensions. For example, College A was judged as more demanding than College C in terms of abstractness, but slightly less demanding than College C in terms of resource use. Additionally, it is of note that there was a clearer difference between colleges' assessments in terms of certain types of demands. For example, the analyses suggest that tasks varied more in demands in relation to complexity than in other respects.

When the analyses were conducted separately for each of the criteria, the separation indices were as follows: 3.75 on complexity (reliability of 0.93), 2.58 on resources (reliability of 0.87), 2.26 on abstractness (reliability of 0.84), 1.42 on response strategy (reliability of 0.67), and 1.79 on task strategy (reliability of 0.76). The separation index for the overall demand was 2.96 (reliability of 0.90). This means that we can be surer that the differences are meaningful for complexity and 'overall demand' than for the other dimensions but the others may also be fair indications of level of demand. Saying this, it should be noted that most of the differences are not particularly large, with most demand values falling between -1 and +1 logits.

In interpreting these results it should also be noted that tutors can take the level of demands of the task into account when making their judgements, and that moderation procedures are in place which should bring college standards in line with national standards where necessary. Evidence of the latter was seen in the study as some of the students at College E failed at the moderation stage, presumably because they had not shown evidence of meeting all the assessment criteria. The lower demands of the tasks at this college could explain why criteria were not met fully (although there are also other possible explanations, such as tasks not being appropriate to test all of the relevant criteria).

Findings from interviews

Interviews were very revealing with findings which included:

- There were substantial similarities between colleges in the tasks conducted.
- There were some differences between colleges that might affect demands, such as the provision of textbook and exemplar reports in some colleges but not in others, and possible differences in the extent to which students must organise their own teamwork.
- There were differences in the extent of use of authentic rather than simulated tasks.
- There were no apparent differences in how tutors made judgements (e.g. they used similar types of evidence).
- It was apparent that some tutors aim for their students to be prepared for the work place and are not just focussed on getting their students to meet minimum standards to pass the assessments.

Discussion of methodology

On a pragmatic level, the methodology seemed to work well. The judges completed all initial comparisons, made suggestions and were able to make paired judgements in most cases. However, some judges did find it difficult initially to get to grips with the notion of 'demands' and some reported that judgements were sometimes difficult.

The judges in the study above each completed an evaluation form about the methodology. Some of the key points arising from this were:

- Most judges agreed that the activities before the meeting were successful in familiarising them with the assessment materials and focussing their thinking on demands.
- The pre-existing framework of demands was felt to provide a useful starting point by most judges and most felt this was preferable to developing a framework from scratch.
- Most judges reported that the revised framework of demands captured all or most of the types of demands placed on students by the assessments.
- There was a feeling from most judges that making judgements was difficult due to assessments often being similarly demanding. 'Abstractness' was considered particularly difficult to judge for this reason.
- The partial nature of the information on the assessment tasks was sometimes a frustration for judges.
- There were a few comments that judging the assessments for the two units separately rather than together would have avoided instances where higher demands in one unit might be averaged against lower demands in the other unit when making the comparative judgements.

These reports suggest that the method was generally successful. It was possible to tailor the framework of demands satisfactorily to the vocationally-related assessment tasks under investigation.

The paired comparisons allowed differences in the profiles of demands between colleges to be identified, and the Rasch analyses gave some impression of the size of these differences. It is of significance that this methodology allows us to produce measures of demand that are amenable to statistical analysis. Additionally, the interview data provided insights into the assessment contexts and into tutors' perceptions in relation to the demands and standards of their assessments and their assessment judgments.

A persisting difficulty with research into demands relates to how we define 'demands' and whether it is possible for judges to make their judgements based on underpinning demands, without being influenced by surface features of task presentation, for example. In the current research we hoped that the initial definition and information, and the initial phase of work would help to ensure a shared and appropriate understanding, but we cannot be entirely sure. This suggests a need for triangulation of such research using additional methods (e.g. using statistical methods that make use of candidates' background attainment).

Perhaps the most problematic element of the methodology is the need for sufficient information on the college-based assessment tasks to be available to judges. It might be possible to improve the methodology by conducting interviews at each of the colleges in advance of the rest of the study and to provide judges with summaries of the information gathered. Similarly, a questionnaire completed by the colleges or training providers in advance might provide appropriate contextual details. Either of these methods would rely on having a reasonable lead-in time to the study in order to gather such information.

Other feedback from the study described above might suggest that it would be better for judgements to be made about each assessment unit separately. It was also possible that the demand dimension 'abstractness' might not be useful for certain assessments and could be excluded. Additionally, more substantial revision of the

starting framework, or for use of an alternative framework developed through the construct elicitation phase, would be possible within this methodology if considered necessary for different types of assessments.

The methodology was generally successful and may provide a useful model for further research on the equivalency of demands of vocationally-orientated assessments across different colleges, training providers or assessment contexts.

References

Adams, R. & Pinot de Moira, A. (2000) *A comparability study in GCSE French, A study based on the Summer 1999 examinations. Review of question paper demand, cross-moderation study and statistical analysis of results.* Organised by WJEC and AQA on behalf of the Joint Forum for the GCSE and GCE.

Arlett, S. (2002) *A Comparability Study in VCE Health and Social Care, Units 1, 2 and 5*. A study based on the Summer 2001 Examination and organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Beaumont, G. (1995) *Review of 100 NVQs and SVQs,* London: National Council for Vocational Qualifications.

Benett, Y. (1993) The validity and reliability of assessments and self-assessments of work-based learning, *Assessment and Evaluation in Higher Education, 18*(2), 83-94.

Boreham, N., Samurçay, R. & Fischer, M. (2002) *Work process knowledge,* London: Routledge.

Crisp, V. & Novakovic, N. (in press) Are all assessments equal? The comparability of demands of college-based assessments in a vocationally-related qualification. *Research in Post Compulsory Education*.

Edwards, E. & Adams, R. (2002) *A Comparability Study in GCE AS Geography including parts of the Scottish Higher Grade examination.* A study based on the Summer 2001 Examination and organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Edwards, E. & Adams, R. (2003) A comparability study in GCE Advanced Level Geography: A review of the examinations requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.

Edwards, J. & Dall'Alba, G. (1981) Development of a scale of cognitive demand for analysis of printed secondary science materials, *Research in Science Education, 11*, 158-170.

Fearnley, A. J. (2000) *A Comparability Study in GCSE Mathematics: a review of the examination requirements and a report on the cross moderation exercise. AQA*.

Gray, E. (1995) *A comparability study in GCSE English, A study based on the Summer 1994 examinations*. Organised by MEG on behalf of the Inter-Group Research Committee for the GCSE.

Gray, E. (2000) *A Comparability Study in GCSE Double Science. A study based on the Summer 1998 examination*. Organised by OCR on behalf of the Joint Forum for GCSE and GCE.

Greatorex, J., Elliott, G. & Bell, J.F. (2002) *A Comparability Study in GCE AS Chemistry. A study based on the Summer 2001 examination.* Organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.

Guthrie, K. (2003) *A Comparability Study in GCE Business Studies, Units 4, 5 and 6 VCE Business, Units 4, 5 and 6*. A study based on the Summer 2002 Examination and organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Houston, J.G. (1981) *Report of the inter-board Cross-moderation Study in Economics at Advanced level: 1979*. Organised by the AEB.

Hughes, S., Pollitt, A. & Ahmed, A. (1998) *The development of a tool for gauging the demands of GCSE and A Level exam questions.* A paper presented at the British Educational Research Association Annual Conference, Belfast.

Jones, B.E. (1993) *GCSE Inter-Group Cross-Moderation Studies 1992*. Summary report on studies undertaken on the Summer 1992 examinations in English, Mathematics and Science.

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

Linacre, J.M. (1989) *Many-faceted Rasch measurement,* Chicago: MESA Press.

Linacre, J.M. (2007) *A User's Guide to FACETS, Rasch-Model Computer Programs, Program Manual*.

Massey, A.J. (1979) *Comparing standards in English Language: A report of the cross-moderation study based on the 1978 Ordinary level examinations of the nine GCE boards.* Southern Universities' Joint Board and Test Development and Research Unit.

McLone, R.R. & Patrick, H. (1990) *A study of the demands made by the two approaches to 'double Mathematics'*. An investigation conducted by the University of Cambridge Local Examinations Syndicate on behalf of the Standing Research Advisory Committee of the GCE Examining Boards.

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.

Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (2007) (Eds) *Techniques for monitoring the comparability of examination standards.* London: QCA.

Oates, T. (1992) Core skills and transfer: aiming high, *Education and Training Technology International, 29*(3), 227-239.

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

Phillips, E. & Adams, R. (1995) *A comparability study in GCSE mathematics. A study based on the Summer 1994 examinations.* WJEC on behalf of the Inter-group Research Committee for the GCSE.

Pollitt, A., Ahmed, A. & Crisp, V. (2007) The demands of exam syllabuses and question papers, in: P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards.* London: QCA.

Pollitt, A., Hutchinson, C., Entwhistle, N. & de Luca, C. (1985) *What makes examination questions difficult?* Edinburgh: Scottish Academic Press.

Qualifications and Curriculum Authority. (2004) *The statutory regulation of external qualifications in England, Wales and Northern Ireland.* London: QCA.

Qualifications and Curriculum Authority. (2008) *Inter-subject comparability studies.* London: QCA.

Ratcliffe, P. (1994) *A comparability study in GCSE Geography. A study based on the Summer 1993 examinations*. Organised by the NEAB on behalf of the Inter-group Committee for the GCSE.

Stobart, G., Elwood, J., Jani, A. & Quinlan, M. (1994) *A comparability study in GCSE History: A study based on the summer 1993 examinations.* Organised by ULEAC on behalf of the Inter-Group Research Committee for the GCSE.

Thurstone, L. L. (1927) A law of comparative judgement, *Psychological Review, 34*, 273-286.

Tolley, H., Greatbatch, D., Bolton, J. & Warmington, P. (2003) *Improving occupational learning: the validity and transferability of NVQs in the workplace,* Sheffield: DfES.

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

Wolf, A. & Silver, R. (1990) *Work based learning: trainee assessment by supervisors.* Manpower Services Commission, Research & Development: No. 33.