Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method

Nadežda Novaković and Jackie Greatorex Research Division

This article is based on a presentation given at The Journal of Vocational Education and Training 8th International Conference held in Worcester College Oxford, UK, in July 2009. The paper was written at the beginning of a wider research project, conducted within the Research Division. The aim of the wider project is to develop a research instrument for comparing the syllabus demands of cognate units from different types of qualifications. The specific aims of the present article are to review the theoretical approaches, methods and research instruments used to compare vocational qualifications (VQs) in England, with the view to gauging their appropriateness for comparing the demands of different types of qualifications. The wider project is still work in progress.

Abstract

Our literature review considers the methods used in studies comparing the demands of vocational syllabus content in England. Generally, categories of demands are either derived from subject experts' views or devised by researchers. Subsequently, subject experts rate each syllabus on each demand category and comparisons can be made. However, problems with the methods include:

- Some studies over-focus on the cognitive domain rather than the affective, interpersonal and psychomotor domains.
- Experts vary in their interpretations of rating scales.

Therefore, we suggest creating a framework of demands which includes all four domains, based on a variety of subject experts' views of demands. The subject experts might rank each syllabus on each type of demand, thus avoiding the problem(s) of rating scales, and facilitating comparisons between syllabuses.

Introduction

Comparability is a complex area of research and investigation, which has been very prominent in the debate about the quality of summative[1] assessment in England in the past decade. This activity has been fuelled by public expectation that assessment standards should remain constant over time, across subjects, between awarding bodies, between test and task demands and so on.

We were tasked with considering methods for comparing the demands of cognate qualifications including vocational qualifications (VQs)[2] in a situation where performance data and performance evidence were lacking and there was limited access to the assessment tasks. This would result in small and/or unrepresentative samples of performance data, performance evidence and assessment tasks. Given the complexity of comparability research we focus on one aspect of comparability: the demands of different qualifications' syllabus content. There are various definitions of 'syllabus', see Nunan (1988) for a detailed discussion. Here syllabus refers to: the statement of the aims/objectives/purpose of the qualification; what knowledge and skills can be in the summative assessment(s); how this will be assessed; and descriptions of levels of quality of performance (e.g. pass or particular grades).

This article presents a review of the relevant research literature relating to comparability within, or at least partly covering, the context of VQs. Different theoretical approaches, methods and research instruments are discussed with the view to gauging their appropriateness for comparing the demands of different types of qualifications.

Comparability of vocational qualifications

A recent publication on the comparability of assessment standards (Newton *et al*., 2007), contains an appendix of 154 comparability reports. However, only seven of these include a VQ, the remainder relate to general qualifications (GQs)[3], illustrating the disparity between comparability research in VQs and GQs.

This disparity is unsurprising, as researching the comparability of VQs is beset by issues not present in the context of GQs. Johnson (2008) indicates that VQs have lower *assessment density* than GQs, *assessment density* refers to the frequency with which assessors judge the same type of performance evidence in similar contexts. Unlike GCSEs[4] and GCE A levels[5], which are mostly assessed through large-scale examinations, VQs tend to be individualised and partly or wholly assessed by criterion-referenced, outcome-based assessment. Therefore, VQ assessors tend to be assessing each candidate's skills and competence based on the evidence of how they perform on specific

[1] Summative assessment is generally used to provide an overall grade or level of achievement for a particular learning programme. Normally summative assessment is used for the purpose of determining who will be awarded a qualification.

[2] Vocational qualifications are designed to focus on learning practical skills (OCR, 2009). Vocationally-related qualifications, give a broad introduction to a particular sector, for example the media or health. For the purpose of brevity we use *vocational qualifications (VQs)* to refer to vocational and vocationally related qualifications.

[3] General qualifications always include examinations as part of the summative assessment. They are broad in nature rather than focused on any particular work-related area (OCR, 2009).

[4] General Certificate of Secondary Education (GCSE). These qualifications are generally taken by 16 year olds at the end of compulsory schooling. Usually students take GCSEs in a series of school subjects. GCSEs are general qualifications.

[5] General Certificate of Education Advanced level qualifications. Generally they are divided into an AS qualification, most often taken by 17 year olds, and A2 assessments, mostly taken by 18 year olds. Combined together the results of the AS and A2 assessments give A level results. A levels are general qualifications.

tasks in specific settings. Candidates' skills may be assessed by a broad range of assessments which may vary considerably from one centre to another (for example, the choice between simulated and authentic activities to assess the same skill).

Most comparability studies focussing on a VQ compare it with a GQ. Greatorex (2001) argues that such comparisons might not be robust due to differences between the qualifications which cannot be accounted for by experimental or statistical controls. For instance: different purposes, learner populations, modes of assessment (e.g. examinations, portfolios) and approaches to applying assessment criteria (e.g. compensation, hurdles).

All the above-mentioned factors might have contributed to the relative paucity of research into the comparability of VQs. However, researching issues relating to comparability in VQs is important for several reasons. First, such investigations are likely to help ensure that VQs are perceived as robust qualifications with consistent standards. Some studies have already been carried out to this effect. For instance, Arlett (2002, 2003) conducted two studies comparing the performance standards and demands of VQs across different awarding bodies in the context of VCE[6] Health and Social Care, a new qualification at the time, and found few large differences. However, Arlett (2003) found a perceived difference in the demand of questions. Guthrie (2003) carried out a similar study comparing GCE Business studies and VCE Business. In many ways the examination and syllabus demands of the VCE versus the GCE were found to be similar. However, the differences in demand between the different types of qualifications were:

- GCE syllabuses encouraged a more synoptic approach than the VCE syllabuses.
- VCE syllabuses encouraged the acquisition of Business skills much more than the GCE syllabuses.
- GCE timed examinations were considered more demanding than the VCE timed examinations.

The research into the comparability between GQs and VQs should also go some way to addressing the dilemmas experienced by employers faced by candidates in possession of different awards – are such qualifications of the same standard, what is the common standard that they share, and what exactly are the differences between them? This is also important as some VQs offer an alternative route to higher education. If two entrants for the same university course both fulfil the requirements for gaining a place but one is in possession of A level qualifications while the other has a VQ, the expectation is that these qualifications should share the same standard.

According to McEwen *et al*. (2001) the traditional view of academic qualifications is that they promote deep conceptual understanding, but may lead to superficial understanding, regurgitation for assessment, and knowledge which cannot be applied outside the narrow range of contexts. On the other hand, GNVQs[7] aimed to integrate 'knowing what' and 'knowing how', but students may not be sufficiently exposed to a wide range of conceptual enquiry and cognitive skills might be neglected (McEwen *et al*., 2001). This is linked to the view that VQs are often seen as an 'easier option' to A levels for lower ability students. According to Barry (1997, p. 44) GNVQs received a lot of "bad press" through some academics condoning the GNVQs as a "second rate" alternative to A levels, and suggesting that the skills learners developed during a GNVQ are gone within a year leaving the learners ill-equipped to study for a single honours degree. Defining the shared standard and clearly stating differences between GQs and VQs might help to address some of the preconceptions currently surrounding VQs.

However, it is unclear how the equivalence of standards should be investigated in such complex cases involving assessments of different nature, designed for different populations of students. Pollitt *et al*. (2007) suggest that no definition of comparability should necessarily be assumed when comparing different assessments and that comparable assessments should not be expected to show the same level in every aspect of demand. Rather, the research should focus on investigating how different demands and different levels of demand present in different assessments balance each other out. "It is asking a lot of examiners to guarantee this balance, and a less ambitious approach requires only that the differences are made clear to everyone involved" (Pollitt *et al*., 2007, p. 166).

In this article, we give an overview of how different studies have approached the task of comparing the demands of VQs, what theories they drew from and which research methods they used to make comparisons. These studies and issues are summarised in Table 1.

Defining demand

Pollitt *et al*. (2007) define 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/ assessments" (2007, p. 196). They are inherent in the assessment tasks and are determined and built into the assessment task during the task writing process. This definition of *demands* makes them distinct from *difficulty*, which refers to how well students perform on an assessment task. While an examination question, for example, may be intended to place little demand on students, and appears to be so to the experts, in reality students may perform poorly due to some question feature overlooked by the question setter. Difficulty can be measured using performance evidence and statistics; demands can be measured only using expert judgement.

Pollitt *et al*.'s definition of demands refers primarily to the assessment task. But many studies, including awarding body studies, have taken a broader definition of demands. In many awarding body studies comparing the demands of examination question papers, mark schemes and syllabus content was a prerequisite to the comparison of performance standards. Examples can be found in the appendix of Newton *et al*. (2007). However, a purely descriptive approach, aiming only to describe various demands, "teachers – even students – might use it when choosing which qualifications to enter for, and employers […] might use it to understand what to expect of those who have taken the exam" (Pollitt *et al*., 2007, p. 167). This type of study is particularly appropriate in situations involving new qualifications. A further step might be to attempt to quantify the relative demand of qualifications using a suitable research instrument(s). In the next sections we consider some of the methods used to describe and compare the demands of VQs.

[6] Vocational Certificate of Education (VCE). This qualification had a similar modular structure to A levels and was principally taken by students of the same age. However, the qualifications were vocational. VCEs are no longer available.

[7] General National Vocational Qualifications (GNVQs) were intended to offer a general introduction to an area of work. They were phased out between 2005 and 2007 (Directgov, 2009).

Table 1: Summary of studies that compare demands and include vocational qualifications

Study	*Qualifications compared*	*Theoretical framework*	*Type of demands compared*	*Focus of study*	*Research instrument*
Barry (1997)	GNVQ Science and A level Chemistry	Marton and Säljös (1976) deep versus surface learning	Teaching and learning styles, content, assessment methods	Curriculum	Participant observation, questionnaires, a test, analysis of relevant documentation
McEwen *et al*. (2001)	GNVQ and A level Science, GNVQ and A level Business	Cognitive development and expertise (Anderson, 1983, Ericcson and Smith, 1991)	Cognitive outcomes	Curriculum	Research (self-observation) diaries
Coles and Matthews (1995)	Various GNVQ and A levels in Biology, Chemistry and Physics	Bloom *et al*. (1956), Gagné (1985), Mitchel and Bartram (1994)	Subject content, general skills, type of performance or learning achievement required by stakeholders, strategies	Summative assessment	Experts identifying the qualification components and skills essential or important for their area of work
SCAA (1995)	Business Studies A level and Advanced GNVQ in Business	No theoretical framework is explicitly provided	Syllabus content, question papers, mark schemes, internal assessment tasks, teaching type and time	Summative assessment	Experts using rating scales on demand categories specified by researchers, interviews
QCA (2006a)	Personal Licence Holder Certificate across different awarding bodies	No theoretical framework is explicitly provided	Cognitive demands, test formats, test content, guided learning hours	Summative assessment	Experts looking for evidence of demand categories specified by researchers
Johnson and Hayward (2008)	Advanced Diplomas (Principal Learning component), BTEC Nationals and A levels in four different contexts: Engineering; IT; Society, Health and Development; Creative and Media	No theoretical framework is explicitly provided	Guided learning hours, content coverage, assessment models, examination requirements	Summative assessment	Experts looking for evidence of demand categories specified by researchers
Arlett (2002, 2003)*	VCE Health and Social Care across different awarding bodies	Personal construct psychology (Kelly, 1955)	Examination question papers, mark schemes, syllabus content, candidates' work	Summative assessment	KRG with rating scales. Rating scales specified by examiners
Guthrie (2003)*	A level Business Studies and VCE Business across different	Personal construct psychology (Kelly, 1955)	Examination question papers, mark schemes, syllabus content, candidates' work	Summative assessment	KRG, rating scales. Rating scales specified by examiners awarding bodies
Crisp and Novaković (2009a, 2009b)	Level 2 Certificate in Administration across centres and over time	Bloom *et al*. (1956), Hughes *et al*. (1998), Kelly's (1995) personal construct psychology	Internally assessed tasks	Summative assessment	CRAS scale, KRG, Thurstone pairs method

Notes *These studies refer to syllabus requirements, which Pollitt *et al*. (2007) refer to as demands, and therefore these studies were included.

Vocational qualification comparability research

The studies comparing VQs included in this review can be divided into two groups.

The first group comprises two studies that have taken a wide view of demands, addressing classroom practices, student learning styles and student cognition in addition to the assessment demands. In this paper, we refer to these studies as focussing on curriculum demands. There are various definitions of 'curriculum', see Nunan (1988) for a detailed discussion. We use 'curriculum' to refer to what is taught, learnt and formatively assessed, the teaching and learning experience, teaching methods, as well as the associated organisation, at the classroom, school and national level. The two studies have drawn on different theories of learning styles and student cognition.

The second group comprise the studies that have focussed primarily on summative assessment demands, such as the demands of examinations, examination questions and tasks, as well as the associated syllabus content. Some studies state that they use Bloom's taxonomy of educational objectives (Bloom *et al*., 1956) and so a short overview of Bloom's taxonomy is provided.

Studies focussing on curriculum demands

Barry (1997) analysed the relative demands of the advanced GNVQ Science and A level Chemistry by comparing the teaching and learning approaches, content and assessment methods associated with each course, using participant observation and questionnaires. GNVQ Science was found to be more conducive to a deep approach to learning than the A level Chemistry course. Furthermore, even though the GNVQ multiple choice questions were considered easier than A level multiple choice questions, in the GNVQ test students had to achieve 70% of marks to pass whilst in the A level test only 40% was required to pass and 70% would constitute a grade A.

McEwen *et al*. (2001) compared A level with GNVQ (in Science and in Business) on three levels: pedagogy, cognitive outcomes and students' metacognition. The authors compared the classroom-based study using self-observation schedules on pedagogy and cognitive outcomes. The authors found a wide overlap in types of learning in the A level and GNVQ classrooms, with some differences. For example, in both A level and GNVQ Science classrooms, there was emphasis on applying theory to practice, problem-solving and developing skills. However, the A level put a lot of focus on memorising, understanding and consolidation, while producing new ideas and being critical were more characteristic of the GNVQ classes. In Business A levels, memorising and consolidation were also reported but more emphasis was placed on student-centred learning than in A level Science. In the Business GNVQ the emphasis was on problem solving and decision making, as well as applying theory to practice.

Studies focussing on summative assessment demands

In the studies focussing on summative assessment demands the choice of specific demand categories was either decided by researchers in advance, or the demand categories were elicited from qualification experts. For the former, researchers drew from an established taxonomy of educational objectives and/or theories of educational and cognitive development, and/or the qualifications under investigation, and/or their experience. Coles and Matthews (1995) is an example of a study that based demand categories or themes on established literature. The method of eliciting demand types on which to compare qualifications from qualification experts was used in three awarding body comparability studies involving VQs (Arlett, 2002, 2003; Guthrie, 2003), and many studies about GQs, the most comprehensive collection of these studies is on the compact disc accompanying Newton *et al*. (2007).

In order to make comparisons Arlett and Guthrie used an initial phase inspired by Kelly's repertory grid (KRG) technique (Kelly, 1955) followed by a comparison of performance standards. The first step involves experts comparing the examination question papers, mark schemes and syllabus content from pairs of qualifications and writing down similarities and differences in demands. These ideas are then discussed and a list of construct statements together with scales for each of these statements is agreed. It is intended that the statements are about demand, and one end of each scale is the least demanding and the other end of the scale is the most demanding. A larger group of expert judges are then asked to rate qualifications on a scale for each of these construct statements. Ratings are usually from 1 to 5 or 1 to 7. Mean ratings can then be compared between syllabuses.

This is a sample of the construct statements from Arlett (2002, p. 3):

> *"Is the question paper layout accessible for candidates?"*
>
> *"To what extent are the questions readable?"*
>
> *"Questions can ask candidates to* recall *information or to* apply *knowledge. What is the relative balance of each in the question papers?"*
>
> *"Is the question structure simple or complex?"*

The studies included a question which asked about the overall level of demand of the syllabus, question papers and mark schemes.

Pollitt *et al*. (2007) argue that one issue with the KRG method is that it generates a wide range of construct statements, some of which do not refer to demands, for example, some are more descriptive, and others refer to how easy it is for the examiner to use the mark scheme. Pollitt *et al*. (2007) suggest that researchers could remove construct statements which do not refer to demands. They argue that the interviews should ask experts to describe *similarities and differences between syllabuses*, and that the interviewers should not steer the interviews to focussing on demand. However, Jankowicz (2004) holds that the interview topic can be determined by the interviewer. Therefore experts could be asked to describe *similarities and differences in demand between syllabuses*, and this might reduce the number of construct statements which are unrelated to demand.

Another method problem highlighted by Pollitt *et al*. (2007) refers to the use of scales, as different judges may apply different values or meanings to the options within the scale. For example, it is quite reasonable to question whether the mid point on a scale represents the same level of demand for a GCE examiner or a GNVQ verifier/moderator, or whether they are basing it on the level of demand of the syllabus with which they are most familiar. Pollitt *et al*. (2007) suggest using a scale from most to least demanding on which the experts rank the syllabuses. Pollitt *et al*.'s suggestion fits with KRG technique as follows. KRG involves two phases, eliciting constructs and then rating or ranking objects on the constructs (Jankowicz, 2004), in our context the objects are syllabuses. An example of a KRG study using ratings is Young *et al*. (2005) and one using rankings is Fransella and Crisp (1979).

Given these method problems, in the following section we consider studies which take a different approach, that is, experts were asked to compare qualifications using a list of demands specified in advance.

Bloom's taxonomy

Bloom's taxonomy (Bloom *et al*, 1956) classifies educational objectives within three domains: cognitive, affective and psychomotor. The taxonomy categories are ordered hierarchically, and are intended to be applicable to all types of education. The taxonomy was designed with several purposes in mind: analysing and developing standards, curricula, teaching and assessment, as well as emphasising alignment between these. It is beyond the scope of this article to discuss alignment, for further information see Maolldomhnaigh and Bealáin (1988), Prophet and Vlaardingerbroek (2003) and Liu and Fulmer (2008).

The cognitive taxonomy is divided into six categories (classes): knowledge, comprehension, application, analysis, synthesis and evaluation. Knowledge (recall of information such as facts or concepts) is the simplest and evaluation (justifying stances by judging the value of information based on a set of criteria) is the most complex. It is beyond the scope of this article to cover the behaviour categories for the affective and psychomotor taxonomies, see Krathwohl *et al*. (1964), Harrow (1972) and Simpson (1972) for details.

In the SCAA[8] (1995) report, subject experts compared the Business Studies GCE and the Advanced GNVQ in Business. They compared the syllabus content, examination question papers, mark schemes and internal assessment tasks on 1) depth and breadth, and 2) skills – *factual recall, planning, investigation, analysis and evaluation, transferability* and *application*, and rated each on a high-medium-low scale. While experts used a rating scale to compare the qualifications, it does not seem they were given examples or guidance as to what would constitute a high or low level of, for example, transferability or recall, highlighting again the problem of using rating scales as a research instrument.

QCA[9] (2006a) reports a study that compared between awarding bodies for the Personal License Holder Certificate[10] by looking into assessment practices across college, employer and training provider centres, as well as the assessment tasks. The study was detailed, covering

[8] SCAA was the School Curriculum and Assessment Authority in England. It was a predecessor of the Qualifications and Curriculum Authority and the Qualifications and Curriculum Development Agency.

[9] Qualifications and Curriculum Authority. The responsibilities of QCA included regulating school examinations in England.

[10] Personal License Holder qualifications are intended for people who will be authorising the supply of alcohol under a Premises licence (QCA, 2006a)

the structure and format of multiple-choice tests, the assessment criteria, mark scheme, demands on candidates and other issues related to the delivery of assessments – maintenance of question item banks, mechanism for issue of results, mechanism for secure delivery, etc. It also made a clear distinction between cognitive demands of the assessment tasks and other types of test demands (text highlighting, option plausibility, reading difficulty, length of options, etc.). QCA (2006b) was a similar study about Door Supervision[11] qualifications.

Regarding the cognitive demands QCA (2006a, 2006b) used a five-level scale, with the levels being: *simple fact recall; complex recall; show understanding of a meaning: simple options; show understanding of a meaning: complex options; and apply reasoning with knowledge* (with *simple fact recall* being the lowest, and *apply reasoning with knowledge* being the highest). In these studies, the experts were not asked to rate the tests on each demand category, but simply state whether there was evidence of any of these in the assessment tasks. If experts found evidence of *simple recall*, that would constitute a demand rating of one, whereas *apply reasoning with knowledge* would constitute a rating of five.

The SCAA and the QCA studies share several features. First, they do not explicitly draw from an established theory or comparability tool. Rather they appear to use a research tool devised by the researchers from their experience. The studies do not provide an indication of the robustness of their research instrument. Secondly, the studies focus primarily on the cognitive domain, whereas the affective and psychomotor domains do not appear to be addressed. Bloom's aim was for educators to focus on all three domains, creating a more holistic form of education. Additionally, many examinations target mostly cognitive outcomes, therefore omitting some important factors, and perhaps distorting educational practice (Martinez, 1999). However, the assessment objectives of some VQs suggest that students should be able to participate in teamwork activities, develop effective communication skills, or effectively perform tasks that involve coordination or physical manipulation of tools. In this sense, any research into the demands of assessment tasks in VQs should take into account the cognitive, affective, interpersonal and/or psychomotor demands, and this has been addressed to some extent by Coles and Matthews (1995, 1998) and Johnson and Hayward (2008).

Coles and Matthews (1995) undertook a comparison of Science GQs and VQs by measuring them against the needs of HE institutions and potential employers. They used a Bloomian model as the starting point, but they adapted it using work by Gagné (1985) and Mitchel and Bartram (1994) to include the skills component, which they termed *practical capability*. The purpose of this was to recognise vocational or applied achievement. The framework they used was thus based around *recall*, *practical capability*, *interpretation*, *application*, *analysis* and *synthesis*. Coles and Matthew's (1995) work was comprehensive[12].

Johnson and Hayward (2008) compared Advanced Diplomas (Principal Learning), BTEC Nationals and A levels. The subject experts rated the requirements for several subjects including Geography, Engineering and Sociology on various issues such as: *knowledge and understanding*, *application and analysis of ideas*, *synthesis and evaluation*, *logical and critical thinking*, *literacy and language skills*, *numeracy skills*, *personal and social skills*, *learning skills*, *vocational and practical skills*. This list appears to focus on the cognitive domain. The purpose of the study was to contribute to the decision of the number of UCAS points each qualification (or each grade that can be awarded for each qualification) was assigned. UCAS points are used in university entrance procedures. Arguably universities are interested in students' cognitive skills which would explain the focus on the cognitive domain. The list above also includes personal and social skills, as well as vocational and practical skills. In this study the experts were required to note the number of times they were able to find evidence of these in the grade descriptors.

Analytic scales of demands

Bloom's taxonomy has partly influenced the development of analytic scales of demands. One such scale (Edwards and Dall'Alba, 1981), was developed in an attempt to quantify the demands placed on secondary school Science students in Australia by lessons, materials and assessments. While drawing on work by Bloom and others, the resulting scale is not a taxonomy. It identifies four categories or dimensions of demand: *complexity*, *openness*, *implicitness and level of abstraction*, and within each of these categories six levels of demand are identified. So, for example, within the complexity dimension the levels progress from *simple* operations (the lowest) to the *evaluation* as the highest. In other words, the entire Bloom's cognitive domain taxonomy is subsumed under only one dimension. The scale was designed to quantify the demands of various subjects. However, a literature search did not reveal any studies using Edwards and Dall'Alba's (1981) scale to compare VQs.

Hughes *et al*. (1998) use Edwards and Dall'Alba's scale as a starting point in developing the CRAS scale of demands. The acronym CRAS refers to the five types of demands contained within the scale:

1) *complexity* (relating to the number of components involved in a task and the relationship between these components);
2) *resources* (relating to the need to use information, either information provided or the student's own internal resources);
3) *abstractness* (the extent to which abstract ideas rather than concrete objects must be used);
4) *task strategy* (the extent to which a strategy for conducting the task must be devised by the student); and
5) *response strategy* (the extent to which a strategy for organising a response must be devised by the student).

The scale contains statements which describe the levels within each dimension, and these can be re-worded for use in different academic subjects. CRAS was developed for summarising the demands of individual assessment tasks. Greatorex and Rushton (2010) compared the CRAS scale with the frames of reference used to compare vocational demands by SCAA (1995), Coles and Matthews (1995, 1998), Arlett (2002, 2003), Guthrie (2003) and QCA (2006a, 2006b). Greatorex and Rushton (2010) concluded that CRAS was too narrow for comparing vocational syllabus demands, because it did not include some of the demands incorporated in the other studies. For instance, Coles and Matthew's (1995) include the demand "more general capabilities such as the ability to work in a team" which is primarily affective and interpersonal, whereas CRAS is predominantly concerned with cognitive demands.

[11] Door supervisors are part of the security teams at public events, public houses etc. Their role includes keeping people safe and checking that only appropriate people enter the venue (Direct.gov.uk, 2010).

[12] Coles and Matthews (1995) compared qualifications in a number of ways, including learning strategies. Arguably, learning strategies are a curriculum rather than a summative assessment issue and therefore this study could be classified as being in the studies comparing qualifications in terms of curriculum demands. However, we classified it as summative assessment demands as this was the focus of their research.

Crisp and Novaković (2009a, 2009b) adapted the CRAS scale for use with vocational assessment tasks by using views from VQ experts generated in a construct elicitation exercise inspired by KRG. Subsequently, the adapted CRAS scale was used to compare the demand of centre-assessed tasks across centres and over time. In order to avoid the previously mentioned problems with using rating scales, the judges were asked not to rate the assessment tasks on each dimension but to make paired comparisons of assessment tasks in terms of each dimension or type of demand. Therefore, for the purposes of this study, the scale was revised to become a framework indicating what makes tasks more or less demanding on each dimension and without a numerical scale, thereby overcoming some of the problems of rating scales.

Crisp and Novaković (2009a, 2009b) also included interviews with some of the centre tutors and students. The results of this strand of enquiry identified some differences between centres that could potentially affect assessment demands. For example, there was some indication that team tasks at one centre placed slightly greater demands on students in terms of organising their group tasks. It was also thought that working with unfamiliar peers rather than friends might alter the demands of team tasks, pointing to the affective task demands. Perhaps the most pertinent difference between tasks related to their degree of authenticity. On one hand the demands of dealing with real events, people or procedures may be higher because the task is more complex and requires more reactive skills. On the other hand, some students may find simulated tasks more demanding in terms of engaging fully with the simulated situation.

The findings of Crisp and Novaković (2009a, 2009b) highlight the need for establishing a framework of demands for VQs that would not focus exclusively on cognitive demands, but would include other types of demands such as interpersonal skills to communicate, interact and influence others to achieve goals with and through others. Studies by Coles and Matthews (1995) and by Johnson and Hayward (2008) acknowledge the limitations of taxonomies focussing only on cognitive demands by adding practical and vocational dimensions to their investigation of comparability. The 'world of work' literature also suggests that extending comparability research beyond the cognitive domain is the right course to follow. For example, McDaniel and Nguyen (2001) report that certain affective factors, such as emotional stability, agreeableness or conscientiousness correlate reasonably well with performance on certain job simulations[13]. Translated to VQs, it is easy to see how learners who have good affective skills may perform well on complex tasks adhering to occupational ethics.

Conclusion

Our review indicates that good practice for studies comparing the syllabus demand of VQs can be summarised as follows:

In the first stage, researchers would conduct KRG interviews with subject experts to elicit demands and statements of what is more and less demanding. This is similar to how many comparability studies have been conducted previously. The aim of this phase is to create a comprehensive framework which will include the cognitive domain as well as the affective, interpersonal and psychomotor domains. To facilitate the inclusion of all domains, at least a section of each interview or some interviews could be devoted to generating demands in each domain. Focussing on each domain was not a feature of many previous comparability studies.

Next, researchers would analyse the constructs into a framework of demands indicating what is more and less demanding. Pollitt *et al*. (2007) suggest that during this process researchers might need to remove some constructs which are not strictly demands.

Following the constitution of a framework several subject experts would rank two or three syllabuses from the most to the least demanding syllabus for each type of demand, thus avoiding the problems of rating scales mentioned previously. Ranking rather than rating is in line with KRG technique, and is suggested by Pollitt *et al*. (2007), but it is a departure from the common use of rating scales. Preferably, the subject experts should rank no more than three syllabuses at a time, otherwise the mental comparisons might become very challenging. The rankings can be used to calculate relative measure of demand for each type of demand.

Given that society's requirement for knowledge and skills often changes, and in turn syllabuses are reworked to reflect these changes, it is likely that any framework of demand would need to be periodically updated.

Final remarks

It was mentioned at the outset that this article was written at the beginning of a wider research project, and the wider project is still work in progress. Since the article was written the research team's thinking has shifted in two ways. First, the present article suggests using only subject experts' views about demands as the basis for a framework of demands for comparing vocational syllabus demand. But current thinking is that both subject experts' views about demands and established research literature should be used to form a framework of demands for comparing syllabus demands of units from different types of qualifications. Second, the research team's thoughts in this article were that subject experts should rank up to three syllabuses rather than use rating scales. The research team's current view is that subject experts should decide which of two units is the most demanding, for several pairs of units. In the wider research literature this process of comparing two items (of any kind) in terms of a particular characteristic is known as the method of paired comparisons. It is a research technique with a long history of use in a variety of contexts including:

- Determining the preferences of preschool children for a series of pictures of play materials (Vance and McCall, 1934).
- Weighting the seriousness of perceived health problems (McKenna *et al*. 1981).
- Comparing the demand of vocational assessment tasks (Crisp and Novaković; 2009a, 2009b).

Despite the changes in the research team's thinking, the present article usefully synthesises literature and makes several timely points.

[13] Situational judgement tests are designed to measure judgement in work settings, and are intended to predict job performance. The tests present test takers with a situation(s) and a list of possible responses. The tests are a form of job simulation. See McDaniel and Nguyen (2001) for further details.

Acknowledgements

Thanks to Hannah Shiell who provided administrative support which facilitated the completion of this article.

References

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press. Cited in A. McEwen, C. McGuinness and D. Knipe. (2001). Teaching and Cognitive Outcomes in A levels and Advanced GNVQs: case studies from science and business classrooms. *Research Papers in Education* **16**, 2, 199–222.

Arlett, S. (2002). *A comparability study in VCE Health and Social Care, Units 1, 2 and 5: A review of the examination requirements and a report on the cross-moderation exercise*. A study based on the Summer 2001 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). A comparability study in VCE Health and Social Care, Units 3, 4 and 6: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Barry, K. (1997). An analysis of the relative demands of advanced GNVQ Science and A level chemistry. *Journal of Further and Higher Education*, **21**, 1, 43–53.

Bloom, B.S., Engelhart, M.D., Furst, E. D., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners*. New York: Longmans.

Coles, M. & Matthews, A. (1995). *Fitness for purpose. A means of comparing qualifications*. London: A report to Sir Ron Dearing.

Coles, M. & Matthews, A. (1998). *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Crisp, V. & Novaković, N. (2009a). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally related qualification. *Research in Post-Compulsory Education*, **14**, 1, 1–18.

Crisp, V. & Novaković, N. (2009b). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation and Research in Education*, **22**, 1, 3–15.

Directgov. (2009). Education and Learning. [online] Available at: http://www.direct.gov.uk/en/EducationAndLearning/QualificationsExplained/DG_10039029 (Accessed 2nd November 2009).

Directgov. (2010). Careers Advice. [online] Available at: http://careersadvice.direct.gov.uk/helpwithyourcareer/jobprofiles/JobProfile?jobprofileid=1242&jobprofilename=Door%20Supervisor&code=-1872239177 (Accessed 30th August 2010).

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.

Ericsson, K.A. & Smith, J. (1991). *Toward a General Theory of Expertise*. Cambridge: Cambridge, University Press. (cited in McEwen *et al.*, 2001).

Fransella, F. & Crisp, A. H. (1979). Comparisons of weight concepts in groups of neurotic, normal and anorexic females, *The British Journal of Psychiatry*, **134**, 79–81.

Gagné, R. M. (1985). The conditions of learning and theory of instruction (4th ed.). New York: Holt, Reinhart and Winston. Cited in M. Coles and A. Matthews (Eds.) 1998. *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Greatorex, J. (2001). *Can vocational A levels be meaningfully compared with other qualifications?* A paper presented at the British Educational Research Association Conference, 13–15 September in Leeds, UK.

Greatorex, J & Rushton, N. (2010). Is CRAS a suitable tool for comparing specification demands from vocational qualifications? *Research Matters: A Cambridge Assessment Publication*, **10**, 40–44.

Guthrie, K. (2003). *A comparability study in GCE Business Studies, Units 4, 5 and 6 VCE Business, Units 4, 5 and: A review of the examination requirements and a report on the cross-moderation exercise*. A study based on the Summer 2002 examination and organised by Edexcel on behalf of the Joint Council for General Qualifications.

Harrow, A. (1972). *A taxonomy of the psychomotor domain: a guide for developing behavioural objectives*. New York: David McKay.

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 conference, September in Belfast, UK.

Jankowicz, D. (2004). *The Easy Guide to Repertory Grids*. Chichester: John Wiley and Sons.

Johnson, J. & Hayward, G. (2008). *Expert group report for award seeking admission to the UCAS tariff: Advanced Diploma*. UCAS. [on line] Available at: http://www.ucas.com/documents/tariff/tariff_reports/adipreport.pdf (Accessed 17th February 2010).

Johnson, M. (2008). Assessing at the borderline: Judging a vocationally related portfolio holistically. *Issues in Educational Research*, **18**, 1. http://www.iier.org.au/iier18/johnson.html (Accessed 30th August 2010)

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

Krathwohl, D.R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of Educational Objectives, Handbook II: Affective domain*. New York: David McKay.

Liu, X. & Fulmer, G. (2008). Alignment between the science curriculum and assessment in selected NY state Regents exams. *Journal of Science Education and Technology*, **17**, 374–383.

Marton, F. & Säljö, R. (1976). On qualitative difference in learning I: Outcome and process. *British Journal of Educational Psychology*, **46**, 4–11.

Maoldomhnaigh, M. Ó. & Bealáin, S. T. Ó. (1988). A comparison of the cognitive demands made by the Integrated Science Curriculum Innovation Project with those made by its written examination for the Intermediate Certificate of Education. *Irish Educational Studies* **7**, 1, 124–133.

Martinez, M.E. (1999). Cognition and the question of test item format. *Educational Psychologist*, **34**, 4, 207–218.

McDaniel, M. & Nguyen. N. (2001). Situational judgment tests: a review of practice and constructs assessed. *International Journal of Selection and Assessment*, **1**, 19–29.

McEwen, A., C. McGuinness & D. Knipe. (2001). Teaching and Cognitive Outcomes in A levels and Advanced GNVQs: case studies from science and business classrooms. *Research Papers in Education*, **16**, 2, 199–222.

McKenna, S., P., Hunt, S, M. & McEwen, J. (1981). Weighting the Seriousness of Perceived Health Problems Using Thurstone's Method of Paired Comparisons. *International Journal of Epidemiology*, **10**, 1, 93–97.

Mitchel. L. & Bartram, D. (1994). The place of knowledge and understanding in the development of the National Vocational Qualifications and Scottish Vocational Qualifications. *Competence and Assessment*, **10**, 1–47. Cited in M. Coles and A. Matthews (eds) 1998. *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

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

Nunan, D. (1988). Syllabus design. In: C.N. Candlin & H.G. Widdowson (Eds.), *Language Teaching: A scheme for teacher education*. Oxford: Oxford University Press.

OCR. (2009). Qualifications. [online] Available at: http://www.ocr.org.uk/qualifications/ (Accessed 2nd November 2009).

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

Prophet, R.B. & Vlaardingerbroek, B. (2003). The relevance of secondary school chemistry education in Botswana: a cognitive development perspective. *International Journal of Educational Development*, **23**: 275–289.

QCA. (2006a). *Comparability study of assessment practice: Personal license holder qualifications*, QCA/06/2709 [online] Available at: http://www.ofqual.gov.uk/files/personal_licence_holder_quals_comparability_study.pdf (Accessed 17th February 2010).

QCA. (2006b). *Comparability study of assessment practice Door supervision qualifications* QCA/06/2710 [online] Available at: http://www.ofqual.gov.uk/files/door_supervision_quals_comparability_report.pdf (Accessed 17th February 2010).

SCAA. (1995). *Report of a comparability exercise into GCE and GNVQ Business*. London: School Curriculum and Assessment Authority.

Simpson E. J. (1972). *The Classification of Educational Objectives in the Psychomotor Domain*. Washington DC: Gryphon House.

Vance, T. F. & McCall, L. T. (1934). Children's Preferences among Play Materials as Determined by the Method of Paired Comparisons of Pictures. *Child Development*, **5**, 3, 267–277.

Young, S.M., Edwards, H.M., McDonald, S. & Thompson, J.B. (2005). Personality Characteristics in an XP Team: A Repertory Grid Study. *SIGSOFT Software Engineering Notes*, **30**, 4, 1–7.

Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work

Tom Bramley Research Division **and Tim Oates** Assessment Research and Development

In this article we describe the method of paired comparisons and its close relative, rank-ordering. Despite early origins, these scaling methods have been introduced into the world of assessment relatively recently, and have the potential to lead to exciting innovations in several aspects of the assessment process. Cambridge Assessment has been at the forefront of these developments and here we summarise the current 'state of play'.

In paired comparison or rank-ordering exercises, experts are asked to place two or more objects into rank order according to some attribute. The 'objects' can be examination scripts, portfolios, individual essays, recordings of oral examinations or musical performances, videos etc; or even examination questions. The attribute is usually 'perceived overall quality', but in the case of examination questions it is 'perceived difficulty'. Analysis of all the judgements creates a scale with each object represented by a number – its 'measure'. The greater the distance between two objects on the scale, the greater the probability that the one with the higher measure would be ranked above the one with the lower measure.

Background

The method of paired comparisons has a long history, originating in the field of psychophysics. Within psychology it is most closely associated with the name of Louis Thurstone, an American psychologist working in the 1920s – 1950s, who showed how the method could be used to scale non-physical, 'subjective' attributes such as 'perceived seriousness of crime', or 'perceived quality of handwriting'.

The method was introduced into examinations research in England in the 1990s principally by Alastair Pollitt, at that time Director of Research at Cambridge Assessment (then known as UCLES – the University of Cambridge Local Examinations Syndicate). He showed how the method could be used for scaling video-recorded performances on speaking tasks in the field of language testing (Pollitt and Murray, 1993), and then went on to apply it to the perennially problematic task of comparing work produced in examinations (in the same subject) from different examination boards, or from different points in time. A detailed description and evaluation of the method's use in 'inter-board comparability studies' can be found in Bramley (2007). Rank ordering is now used extensively in the comparability research work of Cambridge Assessment, and its use in operational aspects of examinations – awarding etc – is being explored and validated. But as with all approaches, it has not and will not be adopted in specific settings without testing its suitability – principally its validity and utility. This requirement for validation is in line with the standards and criteria laid down in The Cambridge Approach.

Although the mathematical details of the method can appear quite complex to non-specialists, at heart the method is very simple, the key idea being that the more times one object 'beats' another in a paired comparison, the further apart they must be on the scale. The resulting scale values are taken to be 'measures' of whatever the comparison was based on, for example 'quality of work produced'. It is assumed that, when comparing work produced in different examinations, the experts making the judgements can allow for any differences in the overall difficulty of the questions or tasks that the examinees were required to respond to.

The main theoretical attraction of the method from the point of view of comparability of examination standards is that the individual judges' personal standards 'cancel out' in the paired comparison method (Andrich, 1978). For example, a judge with a 'severe' personal standard might think that two pieces of work were both worthy of a grade B, while a judge with a more lenient personal standard might think they were both worthy of a grade A – but the two might still agree on which of the pair was better, that is, on the relative ordering of the two pieces of work.