Piloting a method for comparing the demand of vocational qualifications with general qualifications

Jackie Greatorex and **Hannah Shiell** Research Division

Introduction

The demand of vocational and general qualifications receives much attention. One view is that centres offer, and learners take, vocational qualifications as a *purportedly* less demanding route to good grades and to boost centres' performance in league tables (Davis, 2011; Stewart 2010; Paton, 2008, 2010a, b, c). The Wolf Review and a government consultation considered these issues (Wolf, 2011, DfE, 2011a and b). The outcome was that some vocational qualifications remain in league tables (DfE, 2011c). Comparability research helps evaluate the aforementioned view by providing robust information about demand. To undertake such research an instrument is needed. This article reports the pilot of such an instrument. A glossary is provided.

Greatorex and Rushton (2010) and Novaković and Greatorex (2011) reviewed several comparability studies to determine how best to conduct a comparison of demands. All the reviewed studies used a research instrument to gather expert decisions about demands. The reviews indicated that a research instrument should:

- Gather expert decisions in the form of paired comparisons by instructing experts to decide which unit is more demanding. This is repeated for many pairs.
- Cover a variety of domains (areas of knowledge) such as the affective, cognitive and psychomotor domains.

Therefore these two characteristics were incorporated in the instrument piloted here.

The researchers were tasked with investigating methods of comparing general qualifications with vocational qualifications at level 2. Reading a variety of OCR level 2 specifications illustrated that they included knowledge, skills and understanding from five domains (the affective, cognitive, interpersonal, metacognitive and psychomotor domains). Further details about the domains are in Figure 1. These domains were included in the instrument so it should be suitable for use with different types of qualifications.

Domains do not indicate what is more and less demanding; this information is gained from taxonomies. A taxonomy is: "a classification system that establishes the hierarchy of the parts to the whole" (Hauenstein, 1998, 2).

A taxonomy for each domain was chosen from existing literature:

- Affective (Hauenstein, 1998)
- Cognitive (Hauenstein, 1998)
- Interpersonal (Rackham and Morgan, 1977)
- Metacognitive (Howell and Caros, 2006)
- Psychomotor (Hauenstein, 1998).

The demands instrument is included in Appendix A.

Figure 1: Definitions of domains and an example of specification content relating to each domain

Domain of knowledge	*Domain of knowledge relates to:*	*Example of a specification extract relating to the domain*
Affective domain	"developing dispositions (prevailing tendencies) in relation to feelings, values, and beliefs." (Hauenstein, 1998, p.3)	"Knowledge, skills and understanding: Suitable approaches; language and tone; showing empathy." (Certificate in Retail Knowledge, Unit: Understanding customer service in the retail sector, specification p.1)
Cognitive domain	"the process of knowing and development of intellectual skills and abilities." (Hauenstein, 1998, p.3)	"Support conclusions, using reasoned arguments and evidence." (Principal Learning in Engineering, Unit: The engineering world, specification p.35)
Interpersonal domain	Positive relationships between people	"Interact with other speakers and present ideas and information." (GCSE French, Unit: Speaking qualification, specification p.10)
Metacognitive domain	"awareness and conscious use of the psychological processes involved in perception, memory, thinking and learning" (Moseley *et al.*, 2004,p.62).	"The learner can: ... 2.6 Identify own learning needs from feedback obtained from appropriate people." (Certificate in Driving Goods Vehicles, Unit: Make and effective contribution to a business logistics sector, specification. p.2)
Psychomotor domain	"developing physical abilities and skills following an input of information/content." (Hauenstein, 1998, p.3)	"Consistent precision and skill shown in use of apparatus/ equipment. Where appropriate, checks or preliminary work are included to confirm or adapt the apparatus or techniques to ensure data of high quality." (GCSE Chemistry A, Unit: Practical Investigation, specification. p.60)

Comparability studies draw heavily on expert judgement and research about expert judgement has proved fruitful for other areas of assessment practice. For instance, Laming (2004) found that all judgements are relative, that is, they are comparisons of one thing with another. This was used to argue for comparability study methods that ask experts to make relative judgements (paired comparisons/rank ordering). For example, in Kimbell *et al*. (2007) experts were presented with many pairs of scripts and for each pair they decided which script was better. Pairs were constructed from combinations of current and previous exam scripts. The decisions were analysed to determine grade boundaries comparable

with previous grade boundaries. Given that expert judgement research has a track record of being useful, it was decided to further investigate comparability judgement in the present study.

Research questions

Two research questions were investigated:

1. Is the demands instrument appropriate for use in research studies? (i.e. did the results make comparisons between the demand of different types of units?)
2. How did experts judge which units were more demanding?

Judgement questionnaire

A judgement questionnaire was designed to evaluate whether the demands instrument was useable and to investigate how experts judged which units were more demanding. The rationale for each section of the questionnaire is detailed below. The questions from the questionnaire together with the response options are given later in the article.

Instruments are appropriate for use in research if they produce credible results. Previous comparability studies recruited senior assessors, on the basis that their expertise lent credibility to the results and their experience underpinned judgement of demand. (For further details see Elliott and Greatorex (2002) and Adams (2007).) Laming (2004) explains that judgements are heavily influenced by experience. Therefore it is important to know what experience participants thought they used to judge demand. How experience related to judging demand was addressed by Question 1 in the questionnaire.

The experts were instructed to use a concept of typical level 2 learners to judge demand. They cannot follow this instruction if they have no such concept. Therefore Question 2 asked the experts to share their concept of a typical level 2 learner. Additionally, in the interests of transparency it is important to know the basis for judgements.

Research results are invalid if experts judge demand using invalid strategies. Therefore Question 3 investigated judgement strategies, and whether invalid strategies were invoked.

It was expected that experts would find it manageable to hold a concept of typical learners in mind. Novaković (2008) found this to be the case in other assessment situations. Question 4 addressed these issues.

The three anticipated problems of using the demands instrument were that experts were:

1. Judging using concepts other than typical level 2 learners. Experts conceptualised incorrect groups of learners in other assessment situations for example by putting themselves in the place of particular learners (Novaković, 2008) or thinking about familiar learners (Skorupski and Hambleton, 2005).
2. Experiencing difficulty making judgements *using* their concept of typical level 2 learners. Boursicot and Roberts (2006) and Novaković (2008) found evidence of this.
3. Experiencing concept drift, i.e. using different concepts of a typical level 2 learner at different points in the study. Ricker (2006) explained that a limitation of some assessment situations is experts experiencing concept drift.

Questions 5, 6 and 7 addressed these points.

Question 8 explored whether experts agreed with assumptions of research using the demands instrument, such as whether comparisons between specifications from different types of qualifications are meaningful. Question 8 also investigated whether experts agreed that experts in general can do the tasks required by the demands instrument.

Method

Units

Four cognate Health and Social Care level 2 units were selected as listed in Table 1.

Table 1: Type of qualifications and Health and Social Care qualification from which the units were sourced

Type of qualification	*Health and Social Care Qualifications*	*Unit*
VQ	NVQ	NVQ1
GQ	GCSE	GCSE1
VQ	NVQ	NVQ2
GQ	GCSE	GCSE2

The two NVQ units selected were from the same NVQ.

Experts

Four research participants were recruited. They were each a team leader or assistant external verifier or higher for at least one of the qualifications in the study[1] and were all recommended by OCR.

Materials

The experts needed to be familiar with specifications of the units in order to participate in the research. As specifications are substantial documents, which are time-consuming to read, extracts were used rather than whole specifications. The extracts included:

- Aims of the specification
- Assessment objectives of the unit
- Unit content
- Assessment structure
- Information about guided learning hours or assessment time
- Grade/performance descriptors
- Teaching arrangements

The experts were also provided with a document containing the following materials:

- An introduction, instructions, and descriptions of domains and taxonomies
- The demands instrument
- The judgement questionnaire

The demands instrument required experts to compare pairs of units and decide which unit was more demanding for each domain. The experts were also asked to explain their decisions. All possible pairs were compared in this way.

1. This information was taken from OCR records.

Procedure

The experts individually:

- Read the definitions of demands, domain and taxonomy
- Read the specification extracts noting instances of affective, cognitive, interpersonal, metacognitive and psychomotor demands
- Completed the demands instrument and the judgement questionnaire

The experts were sent the materials, which were completed remotely and returned to the Research Division in hard copy.

Analysis

Demands instrument data

The data were analysed in two ways:

1. The level of consensus between experts about whether a unit was more demanding within a domain was calculated. It was noted when all four experts were in consensus that a particular unit in a pair was more demanding in a given domain.
2. The frequency a unit was judged more demanding in a domain was used to rank all four units from the most to the least demanding.

Point 1 focuses on comparing a pair of units, whereas point 2 focuses on comparing all four units.

Judgement questionnaire data

The frequency of responses to closed questions was calculated.

The responses to open questions were divided into sections of text and each section of text was categorised. The frequency of experts whose response was classified in each category was calculated.

The experts' explanations for decisions are not reported here because this is outside the scope of the present study. If there is a consensus amongst experts that there is a difference in demand, the awarding body might decide to change a specification. In such cases the explanations might provide details to guide the changes.

Results

Demands instrument

Consensus amongst experts

Experts individually decided which unit was more demanding for each pair of units. The results are presented in Table 2 which shows the frequency of experts who judged one unit to be the more demanding in a pair for a particular domain. A unit should only be considered to be more demanding when there was a consensus amongst all four experts. The consensus agreements are shaded in the table. For example, four experts judged NVQ1 to be more demanding than GCSE1 in the affective domain, therefore a consensus was reached. Where a cell contains 1, 2 or 3, it indicates there was no consensus on which unit was more demanding.

Table 2 indicates consensus amongst all four experts that:

- An NVQ unit was more demanding than the other NVQ unit in one pair
- A GCSE unit was more demanding than the other GCSE unit in one pair
- An NVQ unit was more demanding than a GCSE unit in six pairs
- A GCSE unit was more demanding than an NVQ unit in three pairs

Comparisons were made between units of the same type, as well as units of different types.

Table 2 Level of consensus between experts

Paired comparison	*Affective*	*Cognitive*	*Interpersonal*	*Metacognitive*	*Psychomotor*
NVQ1 is more demanding than GCSE1	❹	0	❹	1	1
GCSE1 is more demanding than NVQ1	0	❹	0	3	3
NVQ1 is more demanding than NVQ2	2	0	3	2	3
NVQ2 is more demanding than NVQ1	2	❹	1	2	1
NVQ1 is more demanding than GCSE2	2	1	1	3	❹
GCSE2 is more demanding than NVQ1	2	3	3	1	0
GCSE1 is more demanding than NVQ2	0	3	0	❹	3
NVQ2 is more demanding than GCSE1	❹	1	❹	0	1
GCSE1 is more demanding than GCSE2	2	1	2	❹	1
GCSE2 is more demanding than GCSE1	2	3	2	0	3
NVQ2 is more demanding than GCSE2	❹	1	1	3	0
GCSE2 is more demanding than NVQ2	0	3	3	1	❹

❹ indicates all four experts were in consensus that the unit was the more demanding in the pair for a domain

Ranking four units from the most to the least demanding

Table 3 and Figure 2 provide a more holistic picture than that provided in Table 2.

Table 3 and Figure 2 show that for each domain:

- The frequency each unit was judged to be more demanding
- The ranking of units from the most to the least demanding

For instance, in the cognitive domain GCSE2 was judged the most demanding on nine occasions; GCSE1 was judged the most demanding on eight occasions; NVQ2 was judged the most demanding on six occasions; and NVQ1 was judged the most demanding just once. Therefore GCSE2 was ranked the most demanding in the cognitive domain followed by GCSE1, then NVQ2 and NVQ1 was the least demanding.

Whilst it is possible to rank the units in terms of demand the ranks must be treated with caution. Sometimes a unit is ranked higher (or lower) than another unit but the 'differences' in demand are best interpreted as a lack of consensus between experts about whether one unit is the more demanding. Nevertheless, they provide a way of viewing the results for all units at once.

Judgement questionnaire

Table 4 to Table 10 give the frequency of responses to the questions.

Table 4 shows that experts reported that they drew from a variety of experience to make their judgements, particularly GCSE2 experience. For sub-questions *h* to *r* there were some missing responses.

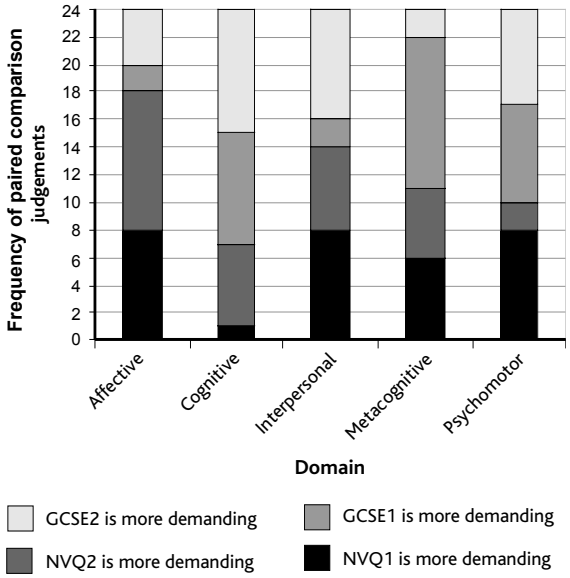

Figure 2: Frequency of paired comparison decisions

Table 3: Frequency that units were judged more demanding, and ranking by demand

Domain	*Unit*	*Frequency unit judged more demanding*	*Rank*
Affective	NVQ2	10	1
	NVQ1	8	2
	GCSE2	4	3
	GCSE1	2	4
Cognitive	GCSE2	9	1
	GCSE1	8	2
	NVQ2	6	3
	NVQ1	1	4
Interpersonal	NVQ1=	8	1=
	GCSE2=	8	1=
	NVQ2	6	3
	GCSE1	2	4
Metacognitive	GCSE1	11	1
	NVQ1	6	2
	NVQ2	5	3
	GCSE2	2	4
Psychomotor	NVQ1	8	1
	GCSE1=	7	2=
	GCSE2=	7	2=
	NVQ2	2	4

Experts generally thought it was easy or very easy to conceptualise particular groups of learners as indicated in Table 5.

Table 6 shows all experts strongly agreed or agreed that they used the concept of typical level 2 learners to decide which unit was more demanding. Fewer experts strongly agreed or agreed they used other concepts of learners to judge demand.

Table 7 shows that experts generally thought it was very easy or easy to use concepts of a group of learners to make judgements. Most importantly, three of the four experts reported it was very easy to use the concept of a typical level 2 learner to make decisions about which unit was the most demanding.

All experts strongly agreed or agreed that they always used the same concept of typical learners as indicated in Table 8. It also shows all experts strongly agreed or agreed they put themselves in the place of familiar GCSE2 learners and thought about what typical GCSE2 learners find more and less demanding. Experts' responses about the other units were more varied. This reflects that experts reported drawing from more GCSE2 than GCSE1 or NVQ experience, as indicated in Table 4.

Table 9 shows all four experts strongly agreed or agreed that specifications from different types of qualifications can be meaningfully compared and writers incorporate the demands they intend learners to experience in specifications. Two experts' responses suggested they thought demand can be judged from specifications, and two experts neither agreed nor disagreed on this issue. Three experts strongly agreed or agreed that:

- For most learners some content is more demanding than other content
- Experts can judge differences in demand between units from different types of qualifications

Table 4: Frequency of responses to question 1

1) How important were the following experiences in your decisions?	*Very important*	*Important*	*Moderately important*	*Of little importance*	*Unimportant*	*N/A*
a. Being a teacher in a subject area relevant to all the units	1	2	0	1	0	0
b. Being a graduate in a subject relevant to all the units	1	1	0	1	1	0
c. Having experience, knowledge and skills in an occupational sector relevant to all the units	2	1	0	0	0	1
d. Being a GQ Chief/Principal/Chair/Assistant Principal Assessor (or equivalent) for GCSE2	2	1	0	1	0	0
e. Being a GQ Team Leader/Assessor (or equivalent) for GCSE2	1	2	0	1	0	0
f. Being a GQ Chief/Principal/Chair/Assistant Principal Assessor (or equivalent) for GCSE1	0	1	0	1	0	2
g. Being a GQ Team Leader/Assessor (or equivalent) for GCSE1	0	1	0	1	0	2
h. Being a NVQ Chief/Principal/Chair/Assistant Principal Assessor (or equivalent) for NVQ2	0	0	0	1	0	2
i. Being a NVQ external verifier for NVQ2	0	0	0	1	0	2
j. Being a NVQ internal verifier for NVQ2	0	0	0	1	0	2
k. Being a NVQ assessor for NVQ2	0	0	0	1	0	2
l. Being a NVQ Chief/Principal/Chair/Assistant Principal Assessor (or equivalent) for NVQ1	0	0	0	1	0	2
m. Being a NVQ external verifier for NVQ1	0	0	1	0	0	2
n. Being a NVQ internal verifier for NVQ1	0	0	1	0	0	2
o. Being a NVQ assessor for NVQ1	1	0	1	0	0	1
p. Having a Level 3 Award Assessing Candidates Using a Range of Methods -A1 (or predecessor awards)	1	0	1	0	0	1
q. Having a Level 4 Award Conducting Internal Quality Assurance of the Assessment Process – V1 (or predecessor awards)	0	0	1	0	0	2
r. Having a Level 4 Award Conducting External Quality Assurance of the Assessment Process	0	0	1	0	0	2

Table 5: Frequency of responses to question 4

4) How easy/hard was it for you to conceptualise the following learners?	Very easy	Easy	Neither easy nor hard	Hard	Very hard	N/A
a. Typical level 2 learners (as described in question 3)	3	0	1	0	0	0
b. Familiar typical level 2 learners	3	1	0	0	0	0
c. The majority of level 2 learners	3	1	0	0	0	0
d. Average level 2 learners	3	1	0	0	0	0
e. Very able level 2 learners	2	2	0	0	0	0
f. Less able level 2 learners	2	2	0	0	0	0

Table 6: Frequency of responses to question 5

5) To what extent do you agree with the following statements? *To decide which unit was more demanding I used my concept of…*	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	N/A
a. Typical level 2 learners (as described in question 3)	2	2	0	0	0	0
b. Familiar typical level 2 learners	1	2	1	0	0	0
c. The majority of level 2 learners	1	2	1	0	0	0
d. Average level 2 learners	1	2	1	0	0	0
e. Very able level 2 learners	1	2	1	0	0	0
f. Less able level 2 learners	1	1	2	0	0	0

Table 7: Frequency of responses to question 6

6) How hard/easy was it for you to use the following concepts to judge which was the most demanding unit using?	Very easy	Easy	Neither easy nor hard	Hard	Very hard	N/A
a. A typical level 2 learner (as described in question 3)	3	0	1	0	0	0
b. Familiar typical level 2 learners	1	2	1	0	0	0
c. The majority of level 2 learners	1	3	0	0	0	0
d. Average level 2 learners	1	1	1	1	0	0
e. Very able level 2 learners	1	2	0	1	0	0
f. Less able level 2 learners	1	1	1	1	0	0

Table 8: Frequency of responses to question 7

7) To what extent do you agree with the following statements?	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	N/A
a. I always used the same concept of typical learners	2	2	0	0	0	0
b. I put myself in the place of familiar GCSE2 learners	2	2	0	0	0	0
c. I thought about what typical GCSE2 learners find more and less demanding	2	2	0	0	0	0
d. I put myself in the place of familiar GCSE1 learners	1	0	2	0	0	1
e. I thought about what typical GCSE1 learners find more and less demanding	1	0	2	0	0	1
f. I put myself in the place of familiar NVQ2 learners	0	0	3	0	0	1
g. I thought about what typical NVQ2 learners find more and less demanding	0	0	3	0	0	1
h. I put myself in the place of familiar NVQ1 learners	0	1	2	0	0	1
i. I thought about what typical NVQ1 learners find more and less demanding	0	1	2	0	0	1
j. I thought about what is more and less demanding for level 2 learners	2	1	1	0	0	0

Table 9: Frequency of responses to question 8

8) To what extent do you agree with the following statements?	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	N/A
a. Specifications from different types of qualifications can be meaningfully compared in a comparability study	2	2	0	0	0	0
b. When specification writers develop specifications they incorporate the demands they intend learners to experience	2	2	0	0	0	0
c. Demands can only be judged from assessment tasks such as exam questions, not specifications	0	0	2	1	1	0
d. Some content of specifications and associated activities is more demanding than others for the majority of learners	2	1	0	1	0	0
e. Experts can rank small numbers of specifications from the most to the least demanding	0	0	4	0	0	0
f. Experts can judge differences in demand between units from different types of qualifications (e.g. general versus vocational qualifications)	2	1	1	0	0	0

Qualitative responses

The experts explained their strategy for making comparisons. There were some strategies which were mentioned by more than two experts such as:

> *Knowledge and understanding required – depth – progression throughout the unit to achieve the intended outcomes*.

However, there were other strategies only mentioned by one expert, such as "Clarity of concepts/instructions."

There are more than four strategies in Table 10 because some experts had more than one approach to making a decision.

Experts described several characteristics of typical level 2 learners, presented in Table 11. Some experts listed more than one characteristic. There were three characteristics each mentioned by two experts, for example, "have breadth but not depth to their knowledge, skills and application". However, there were also an additional nine characteristics and each one was mentioned by only one expert, for example, "Learn by rote".

In Table 10 and Table 11 direct quotes are presented in quotation marks but sometimes the expert views were summarised, in which case quotation marks are not used.

Table 10: Responses to question 2

2) Explain your strategy for deciding which unit was the most demanding	
Example/summary of comments	*Frequency*
"I compared the units side by side looking at the content and the assessment requirements/ guidance for each unit. Alongside looking at the requirements of the domain and how this linked to the skills for each. I then balanced each of the (unit content and domain) to decide which was the most demanding."	2
"Use of particular verbs."	2
"Knowledge and understanding required – depth – progression throughout the unit to achieve the intended outcomes."	2
"Strategies to be used as recommended for assessment purposes."	1
"Range of ways of applying knowledge content to test and recall."	1
"Clarity of concepts/instructions."	1
"Independent research/by candidates overall/holistic approach given to criteria."	1

Table 11: Responses to question 3

3) Please describe your concept of a typical level 2 learner	
Example/summary of comments	*Frequency*
Need help in understanding the specification and its requirements.	2
Have "the basic abilities/ knowledge/skills".	2
Have breadth but not depth to their knowledge, skills and application.	2
"Lack basic knowledge and understanding."	1
Do not use the skills and qualities they already have.	1
"Learn by rote."	1
"Carry out tasks routinely with limited thinking of more complex/varied situations."	1
Are likely to achieve the equivalent of GCSE grade C.	1
Develop their own personal opinions as an independent learner.	1
"Need reassurance before taking on independent research e.g. survey/interviews."	1
"Need clear and concise instructions."	1
"Need look at previous data often not aware of implications of these on future targets."	1

Discussion

The research questions of this study were:

1. Is the demands instrument appropriate for use in research studies?
2. How did experts judge which units were more demanding?

There are several limitations with comparability research in general (see Newton *et al*., 2007). The limitations of the present study include:

- The sample of experts was small (n=4).
- The sample of experts had more GCSE than NVQ experience, according to OCR records. Despite this, where the four experts were in consensus then an NVQ unit was more demanding twice as often as a GCSE unit. Therefore the experts were not biased in favour of their area of experience (GCSE) being more demanding. This is a pleasing result as Massey and Newbould (1977) and Coles and Matthews (1995) found senior examiners judged examinations from their area of experience to be more stringent than other qualifications and Pollitt and Elliott (2003) reported that some subsequent comparability studies were designed accordingly.

It is important to be mindful of these limitations, however, the study does offer useful insights into comparing demands.

1. Is the demands instrument appropriate for use in research studies?

Findings from the demands instrument

The demands instrument is appropriate for use in Cambridge Assessment comparability research of level 2 specifications because the results can be used to compare the demand of units. Evidence for this claim is that in 11 out of 30 pairs of units, there was a consensus between all four experts about which unit was more demanding. A consensus was not expected for all pairs, because units may sometimes be of similar demand.

The results can be used to compare units from different types of qualifications, as intended. For instance, all four experts were in consensus that:

- For six pairs an NVQ unit was more demanding than a GCSE unit
- For three pairs a GCSE unit was more demanding than an NVQ unit

Additionally, the results can be used to compare units from the same type of qualification. For instance, all four experts were in consensus that:

- For one pair, a GCSE unit was more demanding than the other GCSE unit
- For one pair, an NVQ unit was more demanding than the other NVQ unit

Findings from the judgement questionnaire

All the experts agreed or strongly agreed with some of the assumptions of the study. Therefore comparability studies about different types of qualifications have some credibility.

However, one expert did not agree or strongly agree that experts can do the required tasks. This is a cautionary note and suggests that if the demands instrument is used in further research some checks should be made on expert decisions, for example, checking whether one expert consistently disagreed with the others. If so, researchers could consider removing the expert as an outlier. Alternatively, they could check whether the expert panel includes a variety and balance of experience. If not, more decisions may be needed.

2. How did experts judge which units were more demanding?

Responses to the judgement questionnaire

Research instruments are suitable for research purposes if they produce valid results. The validity of results from the research instrument relies on expert judgement. If the experts followed the instructions then the results are valid. If the experts could not or did not follow the instruction then validity was compromised. These principles underpin the discussion below.

Experts' experience

A variety of valid experiences are important in judging, including teaching, qualifications, and experience as an assessor and verifier.

Identifying the concept of a typical level 2 learner

There was some commonality and some individuality in experts' concepts of typical level 2 learners. Diversity can be an advantage if it is more representative. On the other hand, experts sharing a concept can be interpreted as more reliable. The experts generally said it was very easy or easy to conceptualise various groups of level 2 learners, including typical level 2 learners. This was expected because experts find it easy to conceptualise groups of learners in other assessment situations (see Novaković, 2008). It is also encouraging that the experts generally found conceptualising typical level 2 learners manageable, because this is a key part of making judgements and following the instructions.

Using the concept of typical level 2 learners

As instructed, the experts used their concept of typical level 2 learners in their judgements and found it very easy or easy to do so. This finding is positive. However, experts also used and found it easy or very easy to use other concepts of groups of learners such as the "very able" and "less able". It is unclear whether this compromised their judgements. These results were unexpected because in other assessment situations, experts found it difficult to use concepts of groups of learners, for example, Boursicot and Roberts (2006).

Concept drift

All the experts strongly agreed or agreed that they maintained the same concept of typical learners throughout the study. It is pleasing the experts thought they did not experience concept drift which is a limitation of some assessment procedures; for further details see Ricker (2006).

In summary, three problems were anticipated and the questionnaire results suggested only one of these problems occurred; that experts used concepts of non-typical learners. The validity of the research is limited because of this, but overall, the performance of the demands instrument was better than expected.

Conclusion

The demands instrument is considered suitable for use in Cambridge Assessment comparability research for comparing the demand of cognate units in five domains. Such research does not measure the size of differences in demand or the overall demand, and this is important to acknowledge whenever the demands instrument is used. Qualification standards are at specification level. Therefore, a difference in the demand between two units of two different qualifications may not be a cause for concern, so long as the overall demand of the specification is appropriate.

There is little research about the demand of specification content and the pilot suggests that it can be credibly conducted. The comparisons are useful in several areas:

- Qualification development – i.e. checking that draft units are of comparable demand to existing specifications.
- Comparing qualifications when it is not possible to compare learners' performance. This can happen when performance is assessed by observing work-based practice and therefore there are no artefacts created by learners to evidence the quality of performance (Greatorex, 2011).

An area for further research is whether an on screen version of the demands instrument is appropriate for future comparability research at Cambridge Assessment. If so, this would be in line with many examiner activities tending to be undertaken on screen rather than on paper, for example, examination question writing.

Glossary

Cognate "Related or analogous in nature, character, or function." (www.thefreedictionary.com/cognate)

Demanding The extent to which a specification is intended to be challenging for typical learners.

Demand(s) The level of knowledge, skills and understanding required of typical learners to successfully complete a specification. The requirements might be in the: Affective, Cognitive, Interpersonal, Metacognitive and Psychomotor domains.

Demand is a relative term, it could be replaced with 'relative demand' throughout the article. But demand is used for the purposes of brevity.

Domain A domain is a "sphere of knowledge or intellectual activity". (Hauenstein, 1998, 2)

Specification "The complete description – including optional and mandatory aspects – of the content, assessment arrangements and performance requirements for a qualification. A subject specification forms the basis of a course leading to an award or certificate." (www.qcda.org.uk)

Taxonomy A taxonomy is defined as "a classification system that establishes the hierarchy of the parts to the whole. " (Hauenstein, 1998, 2.) Each domain has its own taxonomy. Each taxonomy outlines what is more and less demanding in each domain. The taxonomies and domains are given in.

Unit The smallest part of a qualification for which learners can gain a certificate.

References

Adams, R. (2007). Cross-moderation methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards*. London: QCA.

Boursicot, K. & Roberts, T. (2006). Setting standards in a professional higher education course: Defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school. *Higher Education Quarterly*, **60**, 74–90.

Coles, M. & Matthews, A. (1995). *Report of a comparability exercise into GCE and GNVQ Business*. London: School Curriculum and Assessment Authority.

Davis, A. (2011). Schools to publish results for each subject at GCSE. *London Evening Standard*, 11 March 2011 [online] Available from: http://www.thisislondon.co.uk/standard/article-23931284-schools-to-publish-results-for-each-subject-at-gcse.do accessed 7 June 2011.

Department for Education (DfE) (2011a). Qualifications for 14–16 year olds and Performance Tables. Available from http://www.education.gov.uk/consultations/downloadableDocs/14-16%20policy%20paper%20for%20consultation%20-%20MG%20(3)%20final.pdf accessed on 10 October 2011.

Department for Education (DfE) (2011b). Government publishes response to the Wolf Review of Vocational Education. Available from http://www.education.gov.uk/inthenews/inthenews/a0077253/government-publishes-response-to-the-wolf-review-of-vocational-education accessed on 7 November 2011.

Department for Education (DfE) (2011c). Qualifications for 14–16 Year Olds and Performance Tables. Technical guidance for Awarding Organisations. Available from http://media.education.gov.uk/assets/files/pdf/c/consultation%20response%20on%20qualifications%20for%2014-16-year-olds%20and%20performance%20tables.pdf accessed 28 October 2011.

Elliott, G. & Greatorex, J. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies*, **28**, 253–264.

Greatorex, J. (2011). Comparing different types of qualifications: an alternative comparator. *Research Matters: A Cambridge Assessment Publication*, Special Issue **2**, 3–41.

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.

Hauenstein, A.D. (1998). *A conceptual framework for educational objectives: A holistic approach to traditional taxonomies*. Lanham, MD: University Press of America.

Howell, K. & Caros, J. (2006). Taxonomy of Meta Cognitive Activities: Advanced/Strategic Reading [online] Available at: http://www.wce.wwu.edu/Depts/SPED/Forms/Howell%20-Taxonomy%20%20of%20Strategic%20Reading.pdf. Accessed 16 August 2010.

Kimbell, R., Wheeler, T., Miller, S. & Pollitt, A. (2007). e-scape portfolio assessment: phase 2 report http://www.gold.ac.uk/media/e-scape2.pdf

Laming, D. (2004). *Human judgment: The eye of the beholder*. Thomson: London.

Massey, A. & Newbould, C. (1977). *Comparability by cross moderation: A methodological retreat or a conceptual advance?* A paper prepared for the British Educational Research Association annual conference, Nottingham, UK.

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., Robson, S., Elliott, J. & Gregson, M. (2004). *Thinking skill frameworks for post-16 learners: an evaluation*. A research report for the Learning and Skills Research Centre. Learning and Skills Research Centre, London.

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

Novaković, N. (2008). Thinking on the edge: The influence of discussion and statistical data on awarders' perceptions of borderline candidates in an Angoff awarding meeting. *International Journal of Training Research*, **6**, 1, 74–102.

Novaković, N. & Greatorex, J. (2011). Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method. *Research Matters: A Cambridge Assessment Publication*, **11**, 25–32.

Paton, G. (2008). Vocational courses used to boost results. *The Telegraph*, 20 Aug 2008 [online] Available from: http://www.telegraph.co.uk/education/2592158/Vocational-courses-used-to-boost-results.html Accessed 7 June 2011.

Paton, G. (2010a). Pupils flock to 'less demanding' ICT course. *The Telegraph*, 15 Jan 2010 [online] Available from: http://www.telegraph.co.uk/education/educationnews/6998312/Pupils-flock-to-less-demanding-ICT-course.html Accessed 7 June 2011.

Paton, G. (2010b). GCSE league tables 'skewed by vocational courses'. *The Telegraph*, 15 Feb 2010 Available from http://www.telegraph.co.uk/education/educationnews/7215886/GCSE-league-tables-skewed-by-vocational-courses.html Accessed 7 June 2011.

Paton, G. (2010c). GCSE results: schools use vocational courses to boost scores. *The Telegraph*, 24 Aug 2010 [online] Available from: http://www.telegraph.co.uk/http://www.telegraph.co.uk/education/educationnews/7960154/GCSE-results-schools-use-vocational-courses-to-boost-scores.html Accessed 7 June 2011.

Pollitt, A. & Elliott, G. (2003). Monitoring and investigating comparability: a proper role for human judgement. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability.'

Rackham, N. & Morgan, T. (1977). *Behaviour anaylsis in training*. Maidenhead: McGraw-Hill Book Company.

Ricker, K. (2006). Setting cut-scores: a critical review of the Angoff and modified Angoff methods. *The Alberta Journal of Educational Research*, **52**, 1, 53–64.

Skorupski, W. & Hambleton, R. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, **18**, 3, 233–356.

Stewart, W. (2010). Academies veer towards vocational courses. *TES*, 25 June 2010. [online] Available from: http://www.tes.co.uk/article.aspx?storycode=6048535 accessed 7 June 2011.

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report*. Department for Education. Available from http://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-00031-2011 accessed 27 May 2011

Appendix A: Extracts from the instructions to experts and demand research instrument

Instructions

Enclosed are four specification extracts:

1. NVQ1 2. GCSE1 3. NVQ2 4. GCSE2

Please read and familiarise yourself with the four extracts, and note information in each specification relating to the following five domains of knowledge: affective; cognitive; interpersonal; metacognitive; and psychomotor. The domains are defined below.

If you wish to make notes, please use a separate piece of paper, or write on the specification extracts.

[Figure 1 (*shown on page 29*) was provided here]

Demands research instrument

Introduction

The demands research instrument was developed to compare units in terms of the demand the specification intends to place on typical level 2 learners.

Instructions

- The specification demands research instrument presents you with further information related to the five domains introduced earlier (affective, cognitive, interpersonal, metacognitive and psychomotor). Each domain has its own taxonomy which describes the dispositions and abilities which are more and less demanding for learners within that domain.
- There are several levels within each taxonomy (with the exception of the interpersonal taxonomy). In each case level 1 represents the least demanding learner dispositions/abilities, and each higher level is progressively more demanding for learners. You are **not** required to allocate levels to the specification; the taxonomies and levels are included to aid understanding of the domains.
- Please read the descriptions of the domains and taxonomies and levels on the following pages. Then refer back to your specification extracts and decide which unit intends to place greatest demand on typical level 2 learners in each of the five domains. Use the response sheets to record your decisions.

  Please note that:

  - The most demanding unit may not be the same for all domains.
  - Ties are **not** allowed; you must select one unit as the most demanding for each domain. If you struggled to make a distinction then when you have circled a unit please include a question mark to indicate that you found the judgement difficult.
  - Each comparison must be made independently of the other judgements. In other words, do not deduce the outcome of one comparison from your previous decisions; it is acceptable for your responses to contain inconsistent decisions.
  - You do not need to rely only on the information given explicitly in the specification extracts; some demands may be implicit in the specification, or you might know of them through experience.
  - When making comparisons please remember that the taxonomies describe what is more and less demanding for learners.
  - When making comparisons please remember to concentrate on the demand on typical learners; not more or less able learners, or learners who have special requirements under the Equality Act 2010. (Special considerations are dealt with under other awarding body work.)

You might want to refer to any notes you made on the specifications, about the affective, cognitive, interpersonal, metacognitive and psychomotor.

Descriptions of domains, taxonomies and levels

The next pages present more detailed explanations of the five domains of knowledge (see page [*page number was provided here*] for general domain descriptions and examples from specifications).

The following extended descriptions include domains and their taxonomies for each domain, which describe dispositions and abilities which are more and less demanding for learners. Level 1 represents the least demanding level and the levels then become progressively more demanding. (The interpersonal domain is the exception and does not have levels.) The domains and taxonomies vary in style and structure because they have been developed from different sources.

Please read through the information about each domain and, using this information with the four specification extracts, decide which unit is more demanding. Please record your decision on the response sheets.

Please note that you are not expected to assign taxonomy levels to the specification extracts. The taxonomies are included to give you a better understanding of the domains.

[Adapted versions of the following taxonomies were provided here:

- Affective (Hauenstein, 1998)
- Cognitive (Hauenstein, 1998)
- Interpersonal (Rackham and Morgan, 1977)
- Metacognitive (Howell and Caros, 2006)
- Psychomotor (Hauenstein, 1998).

The affective, cognitive and psychomotor domain levels and descriptions were adapted from Hauenstein, A.D. (1998) *A Conceptual Framework for Educational Objectives, A Holistic Approach to Traditional Taxonomies*. University Press of America: Maryland, with the permission of the publisher.

The interpersonal taxonomy is adapted from Rackham, N. & Morgan, T. (1977) *Behaviour Analysis in Training*, McGraw-Hill: Maidenhead, with permission from Mr N Rackham.

The metacognitive taxonomy is adapted from Howell, K. & Caros, J. (2006) http://www.wce.wwu.edu/Depts/SPED/Forms/Howell%20-Taxonomy%20%20of%20Strategic%20Reading.pdf with the permission of Dr K Howell.

This research instrument including the adapted taxonomies, domains and definitions are only to be used for the purposes of education research by Cambridge Assessment.]

Example Response Sheet

For each row, circle the unit which is more demanding and indicate why the unit is more demanding using the appropriate domain and taxonomy information to explain your decision. If you struggled, include a question mark to indicate that you found the judgement difficult.

Domain	Unit		Why was the more demanding unit more demanding?
Affective	NVQ1	GCSE1	
	NVQ1	NVQ2	
	NVQ1	GCSE2	
	GCSE1	NVQ2	
	GCSE1	GCSE2	
	NVQ2	GCSE2	

The validity of teacher assessed Independent Research Reports contributing to Cambridge Pre-U Global Perspectives and Research

Jackie Greatorex Research Division **and Stuart Shaw** Cambridge International Examinations

Background

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p.9) frame test validity in terms of "the concept or characteristic that a test is designed to measure". That is, the Standards reflect a construct-centred approach to test validity. This perspective draws on the view that the theoretical, underlying construct such as mathematical aptitude, represented by an observable test score is the foundation for evaluating a test. Thus "all test scores are viewed as measures of some construct" (AERA, APA, NCME, 1999, p.174). The claim of validity is that the test adequately reflects the constructs and can be used as basis for the inference of attainment or aptitude depending on the test purpose.

It is important, therefore, to establish that tests elicit performances that reflect intended constructs and that test developers and providers have recourse to a reasonably well-informed and coherent theoretical model underpinning the construct(s) of interest if they are to operationalise aspects of the construct(s) for practical assessment purposes. In reality, however, "Tests are imperfect measures of constructs because they either leave out something that should be included… or else include something that should be left out, or both." (Messick, 1989, p.34). If the construct(s) is not well defined and test tasks are inappropriate, then it will be difficult to support claims an awarding body wishes to make about usefulness of its assessments, including claims that tests do not suffer from construct under-representation and construct irrelevance (CI).

The focus of this research is construct irrelevance. Its working definition for this study is that CI occurs when irrelevant constructs