A method for comparing the demands of specifications

A paper presented at the British Educational Research Association Conference 2012,

and the European Conference on Educational Research 2012

Jackie Greatorex and Sanjana Mehta

Research Division

Cambridge Assessment

Abstract

The comparability of qualifications receives much attention in the UK and there are established methods for conducting post-hoc comparability research. In contrast, this study explored the feasibility of building research into specification development. The aims were to compare the *affective*, *cognitive*, *interpersonal*, *metacognitive* and *psychomotor* demands of new and established units within the same subject, and to identify the relevance of each domain to each unit. Six experts judged which unit (new or established) was the most demanding in each domain. The specifications were qualitatively analysed to explore the relevance of each domain to each unit. As standards reside at qualification level, differences in unit demands may not mean there were differences at qualification level. There were 11 demands-related differences between units. A demands-related difference in a domain which was very relevant to both units was considered more important than a demands-related difference in a domain that was not particularly relevant to the units. For instance, a new unit was more *metacognitively* demanding than an established unit and the *metacognitive* domain was very relevant to both units. So the awarding body might consider decreasing the *metacognitive* demands of the new unit. Demands are complex and judging demands is subjective. Despite these complexities, the study provided credible research evidence about demands. Such studies help to target units in development at the same level of demands as established units. Following the research, the new specification was amended and accredited by Ofqual.

Introduction

Comparability

The UK qualifications system is complex (Wolf, 2011). There are thousands of qualifications of several types and at different levels; for details see Isaacs (2010). (A glossary of terms, common abbreviations and acronyms is given in Appendix 1.) Perhaps this is one reason why the comparability of qualifications receives much attention. One purpose of awarding bodies is to ensure the comparability of various routes to further study and employment.

There are established methods for comparability research (Newton, Baird, Goldstein, Patrick and Tymms, 2007). Such research often compares:

- the demands of assessment assignments e.g. examination questions;
- the quality of learners' performance e.g. as illustrated by their work samples or responses in scripts;
- prior/concurrent measures of attainment.

Examples of comparability studies that focused on one or more of the above are Arlett, (2002, 2003), Coles and Matthews (1995, 1998), Crisp and Novaković (2009a and b), Gray (2000), Fearnley (2000), Greatorex, Elliott and Bell (2002), Greatorex, Hamnett and Bell (2003), Guthrie (2003) and Pritchard, Jani and Monani (2000). Comparability research is usually post-hoc. In contrast, this research explored the feasibility of building research into specification development, by comparing established units and new units in development.

Comparability of demands

It is important to briefly examine the distinction between demands and difficulty. Pollitt, Ahmed and Crisp (2007) reported that demands are the skills required of learners to respond to an assessment assignment. The concept of demands stands in contrast to the concept of difficulty. For Pollitt, Hughes, Ahmed, Fisher-Hoch, and Bramley (1998) difficulty is an empirical measure of how successfully a group of learners performs on an item. In contrast to demands, which have no statistical indicator, difficulty can be explored through statistical techniques such as *facility value*, which:

> is the mean mark on a question expressed as a proportion of the maximum mark available – the lower the facility value the more difficult the question (Pollitt et al., 1998, p.105-106).

Studies about demands have generally focused on the *cognitive* domain (Claisse, 2004; Crisp and Novaković, 2009a; 2009 and b; Emmerich, 1989; Fourali, 1997; Greatorex and Rushton, 2010; Johnson and Mehta, 2011; Ofqual 2011; QCA 2007a; 2007b, 2008a; 2008b; QCA, 2003; Salt, 2005). The influence of such work can be considerable. For instance Johnson and Hayward (2008) and UCAS (2006) analysed the demands of various qualifications to give them a UCAS tariff (point score). The tariff is used to compare qualifications for the purposes of university entry requirements in England and in universities for management information purposes (UCAS 2012a; 2012b; 2012c).

This study took a more holistic view of demands and therefore included five domains: *affective*, *cognitive*, *interpersonal*, *metacognitive* and *psychomotor*. Reading a variety of OCR specifications illustrated that they included knowledge, skills and understanding from these five domains (Greatorex and Shiell, 2012). The demands instrument used in this study included taxonomies to represent the five domains of demands: *affective* (Hauenstein, 1998); *cognitive* (Hauenstein, 1998); *interpersonal* (Rackham and Morgan, 1977); *metacognitive* (Howell and Caros, 2006); and *psychomotor* (Hauenstein, 1998). The taxonomies and domain definitions were adapted with permission for inclusion in the demands instrument. Hereafter "taxonomy" and "domain" will be used instead of "adapted taxonomy" and "adapted

domain" for the purposes of brevity. Greatorex and Shiell (2012) concluded that the demands instrument could be used in research to make credible evidence-based demands comparisons.

Comparing an established specification and a new specification

This comparability study aimed:

- to compare the *affective*, *cognitive*, *interpersonal*, *metacognitive* and *psychomotor* demands of established units and new units;
- to identify the relevance of *affective*, *cognitive*, *interpersonal*, *metacognitive* and *psychomotor* taxonomies to each unit.

The rationale underpinning the second aim was that units should be compared using relevant domains. For instance, if the units tested chairing meetings it was more important to compare the *interpersonal* demands of the units than their *psychomotor* demands. These aims helped explore the feasibility of building research into qualification development. The purpose of the study was also to provide evidence about the demands of the new specification which was independent of evidence from a concurrent Ofqual review. Since the study was completed any necessary amendments have been made to the new specification. The final amended specification was accredited by Ofqual.

The new and established specifications were in the same subject and both were OCR specifications. The specifications were from two different qualifications. The new specification had the following units - N1, N2, N3, N4, N5, N6, N7, N8, N9, N10 and N11. Unit N1 was assessed via an external examination and units N2 to N11 were assessed by controlled assessment. The established specification had the following units – E1, E2, E3, E4 and E5. Units E1 and E3 were assessed using an external examination. E2, E4 and E5 were assessed by controlled assessment. The established specification was taught in centres/schools, but the new specification was in development and as such had never been taught or assessed.

Method

Sample

Seven experts with varied experience in the subject were approached to participate in this study. It was considered that the diversity of the experts' experience would facilitate their judgement processes. The experts, who were recommended by OCR, were invited to participate via an initial email which gave an overview of the study. Individuals who expressed an interest in participating were subsequently sent a more detailed invitation letter and a consent form to sign and return. The experts were sent the materials to make the comparisons. Six experts completed all the activities and provided the data for analysis[1].

Materials

An instrument presenting information relating to five domains (*affective*, *cognitive*, *interpersonal*, *metacognitive*, and *psychomotor*) and taxonomy levels within each domain was provided to the experts. The experts were asked to compare pairs of units and decide which unit was more demanding for typical learners in each of the five domains. They were also asked to give reason(s) for each decision using the appropriate taxonomy.

Extracts from specifications

Specifications are substantial documents and differ in structure. In order to present the most relevant aspects of the specifications and to maintain consistency in presentation, information from the new and established specifications was extracted and these extracts were used in the comparison process. Each extract contained:

- the aims of the specification;
- the assessment objectives of that particular unit;
- the unit content;
- the available grades;
- grade/performance descriptors;
- details about tiering;
- information about guided learning hours and/or assessment time;
- teaching arrangements.

[1] The number of participants with expertise in a qualification was similar to the number in other studies when experts rated demands: that is approximately three. The figure three was calculated from data published in Bramley (2007).

Hereafter the research is described using the terms in Table 1.

Table 1

Terms and definitions

Terms	Description for this study	Description(s) from literature
Duo	A duo of units - one unit was from the established qualification and the other was from the new qualification. Each duo was given a number from 1 to 14. The duos are given in Table 3. The units in each duo were considered to be similar in content.	
Paired comparison	A duo of units compared in terms of what is more and less demanding in a domain.	In general a paired comparison is when two stimuli (A and B) are compared in terms of a criterion (David, 1959; Lee De Simone and Ebrahimi, 2011; Martignon and Hoffrage, 2002). In awarding body comparability literature a paired comparison is often two scripts compared in terms of the quality of the learners' responses to examination questions (Bramley, 2007).
Task	An expert deciding which unit in the paired comparison was more demanding in the domain.	In general a paired comparison task is a participant inferring which object A or B has the higher value on a numerical criterion (Liu, Quinn, Wheeler, Xiao and Lee, 2011; Martignon and Hoffrage, 2002).
Decision	An expert's final judgement about which unit from the task was more demanding within the domain.	
Reason	An expert's justification for their decision.	

Procedure

Each expert was provided with copies of the extracts for the 16 units from the two specifications that were included in this study, and instructions for completing the comparison task. They were instructed to: (i) read the definitions of demands, domain, and taxonomy; (ii) read the extracts noting instances of *affective*, *cognitive*, *interpersonal*, *metacognitive*, and *psychomotor* demands; (iii) complete the demands instrument. Data collection was carried out over approximately three weeks. Each expert completed the activities remotely in their own time. The experts' progress was monitored remotely throughout the time they took to complete the work. Once the experts had completed the activities, they were asked to return all materials, such as specifications, instructions, notes and related material.

The experts were informed that their responses would be used for research purposes, be stored anonymously, and be anonymised in research reports. Each expert was paid a fixed fee for completing all the activities in this study.

Analysis

The data collection aimed to generate one decision per task and a reason for each decision, further details are in Table 2.

Table 2

Structure of a complete data set for six experts

Domain	**Number of paired comparisons**	**Number of experts considering each paired comparison (number of tasks in a paired comparison)**	**Total number of decisions (tasks) per domain**
Affective	14	6	84
Cognitive	14	6	84
Interpersonal	14	6	84
Metacognitive	14	6	84
Psychomotor	14	6	84
Total number of paired comparisons	70		
Total number of decisions (tasks)			420

The frequency of decisions that a unit was the more demanding in a paired comparison was calculated. In Table 3 each cell with a double border contains the two frequencies for a paired comparison. For the purpose of this study, a relative difference in demands between the two units in a paired comparison was assumed if one unit was decided to be more demanding at least 83% of the time (that is, if five or more experts decided a unit was more demanding).

For paired comparisons where five or more experts decided a unit was more demanding, the reasons provided for those decisions were further analysed to determine whether the reasons related to information in the demands taxonomy (in the demands instrument). Taxonomy-related decisions were considered more important and to be indicative of an actual difference in demands as the taxonomies are established and based on research evidence. The decisions which were not taxonomy-related were rejected.

Three members of the research team jointly coded a representative sample of the units from the two specifications to standardise the coding sections and to arrive at a joint understanding of the taxonomies (as given in the demands instrument). Each complete sentence (or main bullet point) was individually coded with between zero and five taxonomies. Following this one team member coded the content of each extract of each unit against each of the five domains. The total number of occurrences of each domain in each unit was noted. Simultaneously, a second member of the research team independently coded one unit from each specification. The codes assigned to these units by the two independent coders were compared for inter-rater reliability.

Results

In this section, the key findings are given in relation to the following results:

1. the frequency of decisions that each unit was more demanding in a paired comparison;
2. analysis of the reasons provided by the experts to determine if their decisions were taxonomy-related;
3. the relevance of the taxonomies to the units;
4. inter-rater coding results.

Decisions and reasons

Each expert decided which unit was more demanding for each task and gave a reason for each decision. When five or more experts decided a unit was more demanding it was considered to indicate a relative difference in demands, for results see Table 3. The relative differences were confirmed upon analysis of the reasons provided for each decision. If the reasons related to taxonomies, then it was concluded that there was an *actual difference* in demands between the units. This further analysis of taxonomy-related reasons is explained later in this section.

Table 3 shows the number of experts who decided a unit was the more demanding in a paired comparison. If five or more experts decided a particular unit in a paired comparison was more demanding then it was most likely not by chance and therefore indicated a relative difference in demands. The reasoning behind this was as follows: assuming the two possible outcomes of a task were equally likely, then the probability of an expert deciding a particular unit was the more demanding was 0.5. A probability of less than 0.05 ($p<0.05$) indicates a rare enough event to accept the research result was not gained by chance. Five or more experts deciding a particular unit in a paired comparison was more demanding was associated with a probability value $p<0.05$.

The units which five or more experts decided were more demanding are indicated in bold in Table 3. For instance, for paired comparison N2 versus E2 in the *metacognitive* domain, five or more experts decided the established unit E2 was more demanding.

These results alone should be interpreted with caution because the differences could be related to a number of variables other than taxonomy-based demands differences. Therefore the reasons for the differences were analysed.

Table 3

Frequency of taxonomy-related decisions for each paired comparison

Duo number	Units in a duo	Affective	Cognitive	Interpersonal	Metacognitive	Psychomotor
1	N1	3	1	2	1	3
	E1	3	5 (4)	4	5 (3)	3
2	N1	2	1	3	1	3
	E3	4	**5 (5)**	3	5 (4)	3
3	N2	2	2	0	1	1
	E2	4	4	**6 (5)**	**5 (5)**	5 (3)
4	N2	3	3	2	3	4
	E3	3	3	4	3	2
5	N3	1	3	0	2	1
	E2	5 (4)	3	**6 (6)**	4	5 (2)
6	N4	1	2	1	2	3
	E2	5 (4)	4	**5 (5)**	4	3
7	N5	1	1	0	2	1
	E4	5 (3)	5 (4)	6 (4)	4	5 (3)
8	N6	1	2	1	2	3
	E4	5 (4)	4	5 (4)	4	3
9	N7	1	2	0	2	3
	E4	5 (3)	4	**6 (5)**	3	3
10	N8	2	2	0	0	2
	E5	4	4	**6 (6)**	5 (4)	4
11	N9	1	2	3	2	3
	E3	5 (3)	4	3	4	3
12	N10	0	0	0	0	3
	E5	5 (1)	**6 (6)**	6 (4)	5 (4)	3
13	N11	2	4	0	**5 (5)**	1
	E2	4	2	**6 (6)**	1 (0)	5 (3)
14	N11	3	2	0	4	1
	E4	3	4	**6 (6)**	2	5 (3)
	Totals	**0**	**2**	**7**	**2**	**0**

Note: The numbers in parentheses indicate the number of decisions that were taxonomy-related.

It can be seen from Table 3 that in the *affective* and the *psychomotor* domains there was no actual difference in demands between the units. In the *cognitive* and *metacognitive* domains, there were two paired comparisons with an actual difference in demands. In the *interpersonal* domain, there were seven paired comparisons for which there was an actual difference in demands. Table 4 lists examples of the taxonomy-related reasons in the *cognitive*, *interpersonal* and *metacognitive* domains.

Table 4

Examples of taxonomy-related reasons provided by the experts for their judgements

Domain	**Decision**	**Example of a reason for the decision**
Cognitive	E3 *is more demanding than* N1	*The use of an unfamiliar context and the need to research a wide range of aspects means the learner must analyse information, break down ideas and make judgements. This is not apparent in the other unit.*
Interpersonal	E2 *is more demanding than* N3	*Learners are required to work with others and as such will be exposed to ideas and proposals from other members of the group which they can support, extend or attack. Similarly they expose themselves to a whole range of responses from others.*
Metacognitive	N11 *is more demanding than* E2	*The project unit requires careful planning and evaluation. The learner must elaborate on the ideas identified by the research making this the most demanding unit in this domain.*

It is encouraging to note that on the whole the missing data was minimal. 416 out of 420 decisions were recorded. There were four missing decisions: two related to duo 12 in the *affective* and *metacognitive* domains and the remainder related to duos 9 and 10 in the *metacognitive* domain.

Relevance of the five domains to the two specifications

Each sentence in each unit was coded against the five domains. Any sentence might be coded with zero to five domains. Table 5 shows the number of times demands in a domain emerged in each unit and across the entire specification. Given the style of coding and that some specifications are more lengthy than others:

- the figures give a broad indication of the relevance of a domain to a unit;
- the row totals differ.

Table 5

Relevance of the five domains (and related taxonomies) to the new and established units

Specification	Unit	Domains				
		affective	cognitive	interpersonal	metacognitive	psychomotor
New	N1	4	34	1	0	3
	N2	0	31	2	0	37
	N3	5	34	0	0	21
	N4	1	24	0	0	21
	N5	0	17	1	0	20
	N6	0	21	3	0	17
	N7	0	19	0	0	19
	N8	0	33	0	0	3
	N9	0	38	1	0	10
	N10	0	28	1	0	12
	N11	0	36	0	19	1
	Total	*10*	*315*	*9*	*19*	*164*
Established	E1	5	66	0	0	5
	E2	2	38	1	9	15
	E3	4	62	0	0	0
	E4	0	33	6	9	18
	E5	0	39	3	6	14
	Total	*11*	*238*	*10*	*24*	*52*

It can be seen that the *cognitive* demands were predominant in both specifications (a total of 315 instances in the new specification and 238 instances in the established specification). This was followed by the *psychomotor* demands (a total of 164 instances in the new specification and 52 instances in established specification). Furthermore, Table 5 illustrates that despite the variability across units, all the five domains occurred in the two specifications, indicating that these demands were relevant to these specifications. Finally, the proportion of occurrence of each of the five demands is also broadly similar across the two specifications. It is also noted that between two and five domains were relevant per unit.

It is not possible to make inter specification comparisons in relation to this relevance mapping due to differences in the length of the units and the number of units in each specification.

Inter-rater reliability was determined by comparing the codes for N7 and E5 which were coded independently by two researchers. Table 6 shows the results of this inter-rater reliability exercise. It can be seen that the two coders reached agreement in most cases. The differences were discussed.

Table 6

Inter-rater coding

		affective	***cognitive***	***interpersonal***	***metacognitive***	***psychomotor***
N7	Coder 1	0	21	0	0	21
	Coder 2	0	19	0	0	19
E5	Coder 1	0	36	2	6	24
	Coder 2	0	39	3	6	14

Discussion

This method of comparing demands is important as it compares demands in terms of evidence based taxonomies and brings comparability research into qualification development. Previously most comparability research was undertaken post-hoc.

Summary of the main findings

In terms of the judgement of demands by the experts, 11 demands related differences were found in the 70 paired comparisons:

- two were related to the *cognitive* domain;
- seven were related to the *interpersonal* domain;
- two were related to the *metacognitive* domain.

The key differences were the two in the *cognitive* domain and one in the *metacognitive* domain. They were important as the *cognitive* and *metacognitive* domains were particularly relevant to the units in these comparisons. Qualification standards are at specification level. Therefore, an actual/key difference in the demands between two units of two different qualifications may not be a cause for concern, so long as the overall demands of the specifications are appropriate.

Comparability research often makes comparisons using constructs that are relevant or similar to the constructs tested by the units/qualifications concerned. For instance comparisons in terms of *cognitive* and *interpersonal* demands might be particularly important when comparing units testing scientific laboratory work conducted in teams, but comparisons in terms of *metacognitive* and *cognitive* demands might be particularly important to comparing units testing individual learners' skills in conducting projects or research work. Therefore an encouraging finding was that the five domains/taxonomies were relevant to the new and established specifications. The *cognitive* and *psychomotor* domains/taxonomies were the most relevant to the specifications. However, not all domains were evident in all units according to the analysis.

Recommendations about specification development

The following recommendations were made to OCR. They illustrate how an awarding body might use such results to modify the demands of units. The recommendations assumed that the established units were of the appropriate demands.

In relation to the taxonomy-related differences OCR might consider prioritising:

- increasing the *cognitive* demands of N1 and N10;
- decreasing the *metacognitive* demands of N11.

These were key differences in demands as the *cognitive* domain was particularly relevant to N1 and N10 and the *metacognitive* domain was particularly relevant to N11. The quantitative research findings did not indicate the scale of the differences in demands. OCR might judge the scale of the differences using the reasons and OCR's expertise in specification development.

In relation to the taxonomy-related differences OCR might also consider increasing:

- the *interpersonal* demands of N2, N3, N4, N7, N8 and N11;
- the *metacognitive* demands of N2.

However, these differences were arguably of less priority as the *interpersonal* domain was not particularly relevant to N2, N3, N4, N7, N8 and N11. Nor was the *metacognitive* domain particularly relevant to N2.

OCR might consider using the reasons to change the demands of the units listed above. The experts' reasons identified the more demanding aspects of a unit. These insights could provide ways of amending the less demanding unit to make the units more explicitly comparable in demands. For example, E3 was judged more demanding than N1 in the *cognitive* domain, one reason given for this was that E3 is more demanding:

> *in "explain ideas etc." and in "extrapolate content information"*

Therefore, to make the demands more comparable, N1 would be amended to more explicitly require learners to explain ideas and extrapolate content information.

OCR could consider whether there are common themes in the reasons which merit particular focus. For example, analysis or analytical skills were mentioned in three reasons for E3 being more demanding than N1. Therefore to make the demands more comparable N1 could be edited to more explicitly require learners to analyse.

Furthermore, if a reason is unclear/inconsistent/irrelevant then OCR might choose not to consider it for specification amendments.

Having received the above recommendations, the Ofqual review, ongoing feedback from colleagues and stakeholder views, OCR decided how to draw from the various sources of information to amend the new specification. The amended new specification was accredited by Ofqual.

Reflections on using the method in comparability research and qualification development

The study had certain limitations, some of which were common to other previous comparability studies. For instance, the experts might have made judgements about the demands in the units based on their subjective and individual understanding of the domains, taxonomies and the examples of demands. Future research might benefit from a standardisation exercise providing more examples of demands from a variety of specifications (which are not being researched) and to discuss them in relation to the domains and taxonomies. Experts should not be standardised on the specifications to be

researched as the rationale for having a variety of experts was to draw from a range of experience and interpretations of those specifications.

The units selected for comparison had suitably similar subject matter to be included in a comparability study, but they had dissimilar grading systems, approaches to controlled assessment and other characteristics. This could have influenced the experts' judgements, although it was not possible to determine the level of influence arising from this aspect.

In terms of methodology, it is acknowledged that the experts carried out the activities remotely and therefore there was limited control over how much time they spent on the activities and the nature and extent of their use of resources. Additionally the sequence of judgements was not part of the instructions or the investigation. This might have affected the comparability judgements given that research shows that in many contexts the order in which judgements are made influences later judgements (Bradburn and Mason, 1964; Kardes and Herr, 1990; Kersholt and Jackson, 1998; Smith, Greenless and Manley 2009). This is not dissimilar to many comparability research activities which are undertaken remotely.

A notable feature of the study was that it compared two specifications of which the established specification was taught and assessed and the new specification had not yet been taught or assessed. Therefore, some of the experts (involved in teaching the established specification) might have had a greater working knowledge of the established specification which in turn may have influenced their judgements. This is different to Greatorex and Shiell (2012) who used the demands instrument with two established qualifications which had both been assessed.

Another aspect of the method is that the experts were asked to compare demands for typical learners. Research shows that people cannot keep a fixed point or model in mind and use it to make consistent judgements (Laming, 2004). For instance there are concerns that experts cannot hold the notion of a typical learner in mind as they make decisions about grade boundaries in the Angoff technique (Impara and Plake, 1997). However, when comparing demands, it was necessary to ask experts to have a particular type of learner in mind, otherwise some might have made comparisons in terms of high achievers and other experts might have made judgements using low achievers. Indeed this might be a strength of the method, as a possibility would be to ask experts to make decisions about demands once concerning A grade learners and once concerning E grade learners. This would provide data about the comparability of demands at each end of the achievement scale. It's interesting to note that experts generally reported it was "very easy" or "easy" to conceptualise various groups of level 2 learners (e.g. typical, very able, less able) and to use their concepts to make judgements about demands (Greatorex and Shiell, 2012).

There are important advantages to this comparability method. The demands instrument includes adapted versions of established taxonomies. Making judgements about demands guided by these taxonomies is likely to be more robust than making judgements without taxonomies. If experts judged demands without the taxonomies, they would have judged the demands of:

- the established specification based solely on their experience of learners who have undertaken the qualification;
- the new specification based on their experience of learners and predicting how they would respond to the new specification.

Another strength of using the demands instrument and associated method was that the demands of new units could be checked against established units. The findings pointed to very targeted areas of the specification which could be changed to adjust demands if necessary. This can be useful in specification development. Other researchers might not have integrated the results about experts' decisions with the relevance of domains to specifications in the manner described above. Future debates might refine how the results are to be integrated.

Conclusion

As noted above, judging demands is complex and involves a good deal of subjectivity. However, comparing the demands of units using taxonomies and expert judgement is a credible comparability method. Such comparisons are useful in:

- specification development – i.e. checking that draft units have comparable demands to existing units;
- comparing specifications or units when it is not possible to use work samples evidencing the quality of learners' performance.

Appendix 1

Glossary of terms used in this paper

Affective	Feelings, values and beliefs (Hauenstein, 1998).
Cognitive	Intellectual skills and abilities (Hauenstein, 1998).
Controlled assessment	"Assessments taken under supervised conditions. They are set by the awarding body and assessed by the learner's teacher or set by the learner's teacher and assessed by an assessor contracted by the awarding body. Many UK qualifications now have controlled assessment rather than coursework" (Greatorex, 2011, 40).
Coursework	"Assessments, often project work, which was set by the learner/teacher/awarding body within awarding body guidelines. Generally assessed by the learner's teacher" (Greatorex, 2011, 40).
Decision	An expert's final judgement about which unit from the task was more demanding within the domain.
Demanding	"The extent to which a specification is intended to be challenging for typical learners" (Greatorex and Shiell, 2012, 35).
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." (Greatorex and Shiell, 2012, 35).
Difficulty	The difficulty of an examination question (or similar) is measured via statistical techniques such as *facility value*, which is the mean mark on a question given as a proportion of the maximum mark available (Pollitt et al., 1998).
Domain	A domain is a "sphere of knowledge or intellectual activity." (Hauenstein, 1998, 2).
Duo	A duo of units - one unit was from the established qualification and the other was from the new qualification. Each duo was given a number from 1 to 14. The units in each duo were considered to be similar in content.
Interpersonal	Relationships between people.
Metacognitive	Consciously using the psychological processes involved in perception, memory, thinking and learning (Moseley et al., 2004).
Psychomotor	Physical skills and abilities (Hauenstein, 1998).
OCR	Oxford Cambridge and RSA. An awarding body.
Ofqual	"National regulator of qualifications in England and vocational qualifications in Northern Ireland" (Greatorex, 2011, 40).

Paired comparison	A duo of units compared in terms of what is more and less demanding in a domain.
QCA	Qualifications and Curriculum Authority. Predecessor of Ofqual.
Qualification level	Qualification levels are within qualification frameworks (e.g. National Qualifications Framework- NQF, Qualifications and Credit Framework - QCF). Each level contains qualifications deemed to be of similar demand. The qualifications in a level vary in subject, content and assessment design.
Reason	An expert's justification for their decision.
Task	An expert deciding which unit in the paired comparison was more demanding in the domain.
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 which outlines what is more and less demanding in each domain.
Type of qualification	Qualifications with a particular characteristic, or from a particular grouping e.g. A-levels, vocational qualifications, BTEC, level 5 qualifications.
UCAS	University and Colleges Admissions Service (UCAS) manages students' applications to higher education (university and college) courses in the UK.
Unit	"The smallest part of a qualification for which learners can gain a certificate" (Greatorex and Shiell, 2012, 35).

References

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 2002 examinations. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

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 examinations. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

Bradburn, N. M., & Mason, W. M. (1964). The effect of question order on responses. *Journal of Marketing Research*, *1* (4), 57-61.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. (pp. 246-294). London: QCA. (CD version).

Claisse, R. (2004). Edexcel Level 4 BTEC Higher Nationals in Computing, Issue 3, B013361, London Qualifications Limited. Retrieved from http://www.edexcel.com/migrationdocuments/BTEC%20Higher%20Nationals/183502_HN_in_Computing_Units.pdf

Coles, M., & Matthews, A. (1995). Fitness for purpose. A means of comparing qualifications. London: A report to Sir Ron Dearing. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

Coles, M., & Matthews, A. (1998). Comparing qualifications – Fitness for purpose. Methodology paper. London: Qualifications and Curriculum Authority. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

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.

David, H. A. (1959). *The method of paired comparisons. A paper presented at the fifth conference on the design of experiments in army research developments and testing*, Fort Detrick, Frederick Maryland http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA417190#page=15

Emmerich, W. (1989). *Appraising the Cognitive Features of Subject Tests*, Research Report No. 89-53, November. Princeton: Educational Testing Service (ETS).

Fearnley, A. (2000). A comparability study in GCSE Mathematics. A review of the examination requirements and a report of the cross moderation exercise. A study based on the Summer 1998 examination and organised by AQA (NEAB) on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

Fourali, C. (1997). Identifying and measuring knowledge in vocational awards: the national vocational qualification experience. *Research in Post-Compulsory Education*, *2* (2), 121-150.

Gray, E. (2000). A comparability study in GCSE Double Science. A review of the examination requirements and a report of the cross moderation exercise. A study based on the Summer 1998 examination and organised by OCR on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

Greatorex, J., (2011). Comparing different types of qualifications: an alternative comparator. *Research Matters: A Cambridge Assessment Publication. Special Issue 2,* 34-41.

Greatorex, J., Elliott, G. & Bell. J. F. (2002). 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 the Research and Evaluation Division, UCLES for OCR on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

Greatorex, J., Hamnett, L. & Bell, J. F. (2003). A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examinations and organised by the Research and Evaluation Division, UCLES for OCR on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

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.

Greatorex, J., & Shiell. H. (2012). Piloting a method for comparing the demand of vocational qualifications with general qualifications. *Research Matters: A Cambridge Assessment Publication. 14,* 29-38.

Guthrie, K. (2003). A comparability study in GCE business studies units 4, 5, and 6 VCE business units 4, 5, and 6. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations. Organised by EdExcel on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

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 Metacognitive activities: Advanced/strategic reading.* Retrieved from http://www.wce.wwu.edu/Depts/SPED/Forms/Howell%20-Taxonomy%20%20of%20Strategic%20Reading.pdf

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34* (4), 353-366.

Isaacs, T. (2010). Educational assessment in England. *Assessment in Education: Principles, policy and practice, 17* (3), 315-334.

Johnson, J., & Hayward, G. (2008). Expert Group report for Award Seeking Admission to the UCAS Tariff. Cambridge Pre-U Diploma. Retrieved from http://www.ucas.co.uk/documents/tariff/tariff_reports/preutariffreport.pdf

Johnson, M., & Mehta, S. (2011). Evaluating the CRAS framework: Development and recommendations. *Research Matters: A Cambridge Assessment Publication*, 12, 27-33.

Kardes, F. R., & Herr, P. M. (1990). Order effects in consumer judgment, choice, and memory: the role of initial processing goals. In M. E. Goldberg, G. Gorn, & R. W. Pollay (Eds) Advances in Consumer Research Volume 17, (pp. 541-546). Provo, UT : Association for Consumer Research.

Kersholt , J, H., & Jackson, J. L. (1998). Judicial decision making: order of evidence presentation and availability of background information. *Applied cognitive psychology*, *12* (5), 445-454.

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

Lee, J-S, De Simone, F., & Ebrahimi, T. (2011). Subjective Quality Evaluation via Paired Comparison: Application to Scalable Video Coding, *IEEE Transactions on multimedia*, *1* (5), 882-893.

Liu, S., Quinn, P.C., Wheeler, A., Xiao, N. Ge, L., & Lee, K. (2011). Similarity and difference in the processing of same- and other-race faces as revealed by eye tracking in 4- to 9-month-olds. *Journal of Experimental Child Psychology*, *108* (1), 180-189.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and decision*, *52* (1), 29-71.

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

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions.* Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. (pp.112-134) London: QCA.

Pritchard, J., Jani, A., & Monani, S. (2000). A comparability study in GCSE English Syllabus review and cross moderation exercise. A study based on the Summer 1998 examination. Organised by Edexcel on behalf of the Joint Council for General Qualifications. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms. (2007) (Eds), *Techniques for monitoring comparability of examination standards*. London: QCA. (CD version).

QCA (2003). Report on Comparability between GCE and International Baccalaureate Examinations. Qualifications and Curriculum Authority. Retrieved from http://www.ofqual.gov.uk/files/alevels_vs_ib.pdf

QCA (2007a). GCSE (short course) citizenship studies comparability study. QCA/07/3089. Qualifications and Curriculum Authority. Retrieved from http://www.ofqual.gov.uk/files/QCA-07-3089_GCSE_citizenship_mar07.pdf

QCA (2007b). GCSE French comparability study. QCA/07/3098. Qualifications and Curriculum Authority. Retrieved from http://www.ofqual.gov.uk/files/QCA-07-3098_GCSE_French_mar07.pdf

QCA (2008a). Inter-subject comparability studies, Study 1a: GCSE, AS and A level geography and history. QCA/08/3651. Qualifications and Curriculum Authority. Retrieved from http://www.ofqual.gov.uk/files/qca-08-3651_study_1a.pdf

QCA (2008b). Inter-subject comparability studies, Study 2b: A level English literature, history and media studies. QCA/08/3654. Qualifications and Curriculum Authority. Retrieved from http://www.ofqual.gov.uk/files/qca-08-3654_study_2b.pdf

Ofqual (2011). Review of Standards in GCE A level English Literature 2005 and 2009 Ofqual/11/4847. Coventry, Belfast: Office of Qualifications and Examinations Regulation. Retrieved from http://www.ofqual.gov.uk/news-and-announcements/83/739

Rackham, N., & Morgan, T. (1977). *Behaviour Analysis in Training*. Maidenhead: McGraw-Hill.

Salt, N. (2005). *A Model IT Curriculum for ESL Students*. Proceedings of the 6th conference on Information technology education of the Special Interest Group Information Technology Education (SIGITE) 2005 Association for Computing Machinery (ACM) New York, NY, USA.

Smith, M. J., Greenless, I., & Manley, A. (2009). Influence of order effects and mode of judgement on assessments of ability in sport. *Journal of Sports Sciences*, *27* (7), 75-752.

UCAS (2006). Expert Group report for Award Seeking Admission to the UCAS Tariff. International Baccalaureate. Retrieved from http://www.ucas.com/documents/tariff/tariff_reports/ibreport.pdf

UCAS (2012a). Thinking about demand. Retrieved from http://www.ucas.ac.uk/documents/qireview/qirdemandpaper.pdf

UCAS (2012b). Qualifications Information Review Briefing. Retrieved from http://www.ucas.ac.uk/documents/qireview/qirbriefing.pdf

UCAS (2012c). Qualifications Information Review Consultation. Cheltenham: UCAS. Retrieved from http://www.ucas.ac.uk/documents/qireview/qirconsultation_english.pdf

Wolf, A. (2011). Review of Vocational Education – The Wolf Report. Department for Education. Retrieved from http://www.education.gov.uk/publications/standard/publicationDetail/page1/DFE-00031-2011