Using scales of cognitive demand in a validation study of Cambridge International A and AS level Economics

Jackie Greatorex Research Division, **Stuart Shaw** CIE, **Phineas Hodson** CIE **and Jo Ireland** Research Division

Introduction

The CRAS (Complexity, Resources, Abstractness and Strategy) framework is used to evaluate the cognitive demands of examination questions (Crisp and Novaković, 2009a, 2009b; Hughes, Pollitt and Ahmed, 1998; Pollitt, Ahmed and Crisp, 2007; Pollitt, Hughes, Ahmed, Fisher-Hoch and Bramley, 1998). Johnson and Mehta (2011) reviewed how CRAS was used; they endorsed some practices and made several recommendations (detailed below). This article provides an illustration of Johnson and Mehta's (2011) principles for using CRAS in the context of validating examination questions used in Cambridge International A and AS level Economics, and highlights some advantages and difficulties inherent in their methods. This article is part of our exploration of how to refine the use of CRAS, particularly with multiple choice question papers, with the aim of sharing issues and recommendations.

Development of CRAS

CRAS was developed from earlier scales of cognitive demands (Edwards and Dall'Alba, 1981) combined with examiners' views about what is more and less demanding for candidates (Hughes *et al*., 1998). Hughes *et al*. (1998) and Pollitt *et al*. (1998) describe CRAS as having four dimensions: Complexity, Resources, Abstractness and Strategy. This was increased to five by Pollitt *et al*. (2007), who split Strategy into Task strategy and Response strategy (Figure 1).

It is Pollitt *et al*.'s (2007) conceptualisation of the dimensions that was used in this study. It has also been used by other researchers, for example, Crisp and Novaković (2009a).

Johnson and Mehta (2011) make several recommendations for using CRAS including the following:

- CRAS should only be used where the CRAS dimensions map to the constructs to be examined. CRAS is predominantly cognitive so it is only suitable for evaluating cognitive demands.
- Individual examination questions may be evaluated using CRAS but ratings on the dimensions may not be summed to give a value for the overall demand of an examination paper.
- An expert's rating of an examination question on one dimension can be compared with their rating of another examination question on the same dimension.
- Ratings on the different dimensions should not be combined to give individual questions a score for 'total demand'.

CRAS has been used to evaluate examination questions but it could also be used to evaluate the cognitive demands of text books, curricula, lesson contexts and marking criteria (Hughes *et al*., 1998; Johnson and Mehta, 2011) and in validation studies (Shaw and Crisp, 2012; Shaw, Crisp and Johnson, 2011).

Figure 1: The CRAS Scales of demands (Pollitt *et al*., 2007, 186)

	1	*2*	*3*	*4*	*5*
Complexity The number of components or operations or ideas and the links between them.		Mostly single ideas or simple steps. Little comprehension, except that required for natural language. Few links between operations.		Synthesis or evaluation is required. Need for technical comprehension. Make links between cognitive operations.	
Resources The use of data and information.		More or less all and only the data/ information needed are given.		Student must generate or select the necessary data/information.	
Abstractness The extent to which the student deals with ideas rather than concrete objects of phenomena.		Mostly deals with concrete objects.		Mostly abstract.	
Task strategy The extent to which the student devises (or selects) and maintains a strategy for tackling the question.		Strategy is given. Little or no need to monitor strategy. Little selection of information required.		Students need to devise their own strategy. Students must monitor the application of their strategy.	
Response strategy The extent to which students have to organise their own response.		Organisation of response hardly required.		Must select answer content from a large pool of possibilities. Must organise how to communicate response.	

Table reproduced from 'Techniques for monitoring the comparability of examination standards', © Qualifications and Curriculum Authority 2007.

Cambridge International A and AS level Economics

Cambridge International Examinations (CIE) A and AS level Economics examinations are offered to students across the world in two series, November and June (CIE, 2009). This article reports on research which uses papers from the June 2011 series. The examination comprises four papers:

- AS level – multiple choice (30 questions)
- AS level – data response and structured essay (4 questions)
- A level – multiple choice (30 questions)
- A level – data response and structured essays (7 questions)

The five Assessment Objectives (AOs) for A and AS level Economics state that students are expected to:

- Demonstrate knowledge and understanding of the specified content.
- Interpret economic information presented in verbal, numerical or graphical form.
- Explain and analyse economic issues and arguments, using relevant economic concepts, theories and information.
- Evaluate economic information, arguments, proposals and policies, taking into consideration relevant information and theory, and distinguishing facts from hypothetical statements and value judgements.
- Organise, present and communicate economic ideas and informed judgements in a clear, logical and appropriate form (CIE, 2009, 5).

Validation

Educational measurement and psychological testing generally take a construct-centred approach to the validity of question papers, psychological tests and other assessments (Brown, 2010; Ertl and Stasz, 2010; Kane, 2009; Messick, 1995; Quinlan, Higgins and Wolff, 2009; Shaw *et al*., 2011; Stobart, 2009; Threlfall, Nelson and Walker, 2007; Tran, Griffin and Nguyen, 2010; Vogt, Proctor, King, King and Vasterling, 2008). A construct-centred approach draws on the view that an underlying theoretical construct, such as mathematical aptitude, is represented by an examination mark and is the foundation on which the evaluation of an examination is built (Messick, 1989). A claim that an interpretation or use is valid must be backed by evidence that the marks from the examination adequately reflect the constructs.

To establish whether examinations elicit performances that reflect intended constructs awarding bodies must have recourse to a reasonably well-informed and coherent theoretical model underpinning the constructs of interest. The work from which this CRAS study was drawn utilised the model for validation of general qualifications proposed by Shaw *et al*. (2011), and illustrated in Shaw and Crisp (2012), which is itself situated in Kane's model of validation through argument (Kane, 1992). Based on this theoretical background, the CRAS framework was used to answer the following validation questions for CIE Economics A level:

- Do the tasks elicit performances that reflect the intended constructs?
- Do the tasks adequately sample the constructs that are set out as important in the syllabus?

The AOs are assumed to represent the intended constructs. In line with Johnson and Mehta's (2011) recommendations, the constructs were broadly mapped to the CRAS dimensions (Figure 2). Additionally, the item types were judged by the researchers to broadly map to the CRAS dimensions, with the caveat that Response strategy was less relevant to multiple choice questions than other question types. Shaw and Crisp (2012) did not report problems applying Response strategy to multiple choice questions, and on the basis of this research evidence we considered all CRAS dimensions to be suitable for use with all the question types.

Method

Six experts applied the CRAS instrument to the selected question papers, two of which contained multiple choice items and two of which contained essay and data response items. The experts were chosen on the basis of their experience as senior examiners for CIE 16–19 Economics qualifications.

Each expert was issued with the following materials:

- task instructions
- a copy of each of the question papers
- a copy of the mark scheme for each question paper
- the CRAS scales
- a response sheet for each question paper (see Appendix A for an example).

The instructions informed the experts that the exercise was about the cognitive demands of examination questions. For Task 1 they were instructed to:

- ignore the mark scheme
- familiarise themselves with CRAS and the question papers
- look at each question on each paper individually
- for each type of demand on each question, rate the level of demand of the activities the students have to do to answer the question on a scale of 1 (low demand) to 5 (high demand)
- work remotely and individually.

For Task 2 they were asked to repeat the exercise focusing on the demands rewarded by the mark scheme.

Analysis

The ratings across the five dimensions given to each question by each expert were tabulated to provide an indication of the range of demands across questions and the level of inter-rater agreement. Ratings without the mark scheme allowed inferences about construct elicitation to be made from the CRAS demands of the questions, that is, they indicated how demanding the question would appear to a candidate, teacher or other stakeholder if they did not consult the mark scheme. Ratings with the mark scheme allowed inferences about whether the mark scheme rewarded (sampled) the constructs. For the purposes of brevity the report refers to the ratings indicating the demands rewarded by the *mark scheme*, which is the primary focus of Task 2, however these ratings indicate the demands rewarded by the combination of the *question paper and the mark scheme*.

To investigate the degree to which the mark scheme rewarded the demands inherent in the question, each expert's ratings with and without the mark scheme were compared. The frequency of experts giving different ratings with and without the mark scheme was calculated for each dimension.

The scope for summary statistics was constrained by the ordinal nature of the data and the principle that ratings from different experts cannot be combined. Based on the sample size in this study most quantification was precluded unless questionable assumptions were introduced concerning equal-interval scales and common internalisation of the scales and anchor points.

Figure 2: Mapping CRAS to the assessment objectives and question types

Dimension description from Pollitt et al. *(2007)*	*Reason(s) for relevance to the AOs*	*Reason(s) for relevance to question types*
Complexity The number of components or operations or ideas involved in a task and the links between them.	The skills in the AOs (demonstrating understanding, interpreting, explaining, analysing, evaluating, organising and communicating) can involve one or more steps, technical comprehension and synthesis or evaluation.	All question types: • Can involve one or more ideas/steps • Relate to technical information • Can involve links between operations (evaluation/synthesis).
Resources The use of data and information.	Using data and information correctly requires knowledge and understanding of economics (AO1). The student clearly generating information involves: • Interpreting information presented in verbal, numerical or graphical form (AO2). • Explaining and analysing (AO3). • Evaluating (AO4). • Taking into consideration relevant information and theory and distinguishing facts (AO4). • Organising, presenting and communicating (AO5).	For multiple choice and data response students are provided with data/information such as text, graphs and statistics and can require: • Using only the data/information provided • Generating data/information. For essays students must generate much of the necessary data/information.
Abstractness The extent to which the students must deal with ideas rather than concrete objects.	All the AOs involve dealing with abstract information. .	For all question types the content is abstract.
Task strategy The extent to which the students must devise (or select) and maintain a strategy for tackling the question.	Strategies might involve any combination of or all of the skills in the AOs.	Questions of all types can involve being given a strategy or devising a strategy and monitoring the application of the strategy.
Response strategy The extent to which the students have to organise their own response.	Reflects AO5 which requires the students to organise, present and communicate economic ideas and informed judgements in a clear, logical and appropriate manner.	The essay questions and to a lesser extent the data response questions require students to organise their own response.
Does CRAS map to the AOs and question types? Complexity, Resources, Abstractness and Task strategy mapped to all the AOs and question types. Response strategy reflected AO5 rather than the other AOs and was more relevant to essay and data response questions than to multiple choice questions. Therefore the constructs broadly mapped to CRAS.		

Validation findings

The findings which follow are based solely on the data response and essay questions. The multiple choice papers are dealt with later under a separate heading.

There was a strong tendency for each rater to place many of the questions on a paper at the same level of demand (see Tables 1 to 4). In some contexts this could suggest a threat to validity as, if all questions are of similar demand, low ability candidates may lack sufficient low-demand questions to demonstrate their abilities and/or high ability candidates may not be stretched by sufficiently demanding questions. However, where candidates choose between optional essay questions, as is the case here, consistent demand is a desirable feature since whatever questions candidates choose, they will experience similar demands. Much greater diversity in demand levels is shown by inter-rater comparisons but, in the absence of evidence that the internalised scales of raters were similar, valid comparisons at this level are not possible.

The comparison between the demands elicited by the questions and the demands rewarded by the mark scheme, as illustrated in Figures 3 to 7, shows that there were no questions for which a consensus, or something approaching a consensus, existed on which was more demanding. As the raters as a group considered demands to be equal, or were divided on whether it was the questions or the mark schemes which embodied greater demand, it can be inferred that demands were broadly similar across the two. This suggests that the demands rewarded by the mark scheme and those elicited by the question were similar, which provides evidence of validity.

This study alone should not be seen as providing a compelling answer to the validation questions, but can provide a valuable perspective and contribute to the body of evidence. The small number of experts used does reduce the power of this study but the number used is sufficient to warrant the conclusions and the research effort required for large sample sizes will not always be available.

Table 1: Expert ratings of questions from question paper 2 without the mark scheme

Demand level (rating category)		*Expert 1*	*Expert 2*	*Expert 3*	*Expert 4*	*Expert 5*	*Expert 6*
Complexity	5 (High demand)	1 2 3 4		2 3 4			1
	4		4	1	1	2 3 4	4
	3		1 2 3		2 3 4	1	
	2						2 3
	1 (Low demand)						
Resources	5 (High demand)	2 3 4				2 3 4	3
	4	1		2 3 4			1 2 4
	3		1 2 3 4		1	1	
	2			1			
	1 (Low demand)				2 3 4		
Abstractness	5 (High demand)	4				2 4	
	4	1 2 3		2 3	1 2 4	1 3	2 4
	3		2 4	4	3		1 3
	2		1 3	1			
	1 (Low demand)						
Task strategy	5 (High demand)	1 2 3 4		1 2 3 4		2 3 4	
	4						2 3 4
	3		1 2 3 4		1 3 4	1	1
	2				2		
	1 (Low demand)						
Response strategy	5 (High demand)			1 2 3 4		2 3 4	1 2 3 4
	4	2 3 4					
	3	1	1 4		1 3 4	1	
	2		2 3		2		
	1 (Low demand)						

Note: Question numbers appear in the cells in columns 3 to 8.

Table 2: Expert ratings of questions from question paper 2 with the mark scheme

Demand level (rating category)		*Expert 1*	*Expert 2*	*Expert 3*	*Expert 4*	*Expert 5*	*Expert 6*
Complexity	5 (High demand)	1 4					1 4
	4	2 3	4	1 2 3 4	1	2 3 4	2 3
	3		1 2		2 3 4	1	
	2		3				
	1 (Low demand)						
Resources	5 (High demand)	2 3 4		4		2 3 4	
	4		2 3	3	1	1	1 2 3 4
	3	1	1 4	2			
	2			1			
	1 (Low demand)				2 3 4		
Abstractness	5 (High demand)	4				2 4	3
	4	1 2 3	2 4	2 3		3	1 2 4
	3		3	4	1 2 3 4	1	
	2			1			
	1 (Low demand)		1				
Task strategy	5 (High demand)	2 3 4		1 2 3 4			1 2 3 4
	4	1			3 4	2 3 4	
	3		1 2 3 4			1	
	2				1 2		
	1 (Low demand)						
Response strategy	5 (High demand)			1 2 3 4		2 3 4	1 2 3 4
	4	2 3 4	1 3				
	3	1	2 4		3 4	1	
	2				1 2		
	1 (Low demand)						

Note: Question numbers appear in the cells in columns 3 to 8.

Table 3: Expert ratings of questions from question paper 4 without the mark scheme

Demand level (rating category)		*Expert 1*	*Expert 2*	*Expert 3*	*Expert 4*	*Expert 5*	*Expert 6*
Complexity	5 (High demand)	1 2 3 4 5 6 7		2 3 4 5 6 7		4 6 7	1
	4		1 2 5 6	1	1	2 3 5	6 7
	3		3 4 7		3 5 6 7	1	3 5
	2				2 4		2 4
	1 (Low demand)						
Resources	5 (High demand)	2 3 4 5 6 7				2 3 4 5 6 7	
	4		1 2 4 5 6 7	2 3 4 5 6 7	1	1	1 2 3 4 5 6 7
	3	1	3				
	2			1	2 3 4 5 6 7		
	1 (Low demand)						
Abstractness	5 (High demand)	2 4 6	2 6			2 6 7	1 4
	4	1 3 5 7	1 4 5	2 3 4 5 6 7	2 3 4 5 6 7	1 3 4 5	2 3 5 6
	3		3 7		1		7
	2			1			
	1 (Low demand)						
Task strategy	5 (High demand)	1 2 3 4 5 6 7		1 2 3 4 5 6 7		3 5 6	1 4
	4		1 2 4 5 6		4 7	2 4 7	2 3 5 6 7
	3		3 7		1 2 3 5 6	1	
	2						
	1 (Low demand)						
Response strategy	5 (High demand)	4 7		1 2 3 4 5 6 7		2 3 4 5 6 7	1 2 3 4 5 6 7
	4	2 3 5 6	4 7		4 7	1	
	3	1	2 3 5		1 2 3 5 6		
	2		1 6				
	1 (Low demand)						

Note: Question numbers appear in the cells in columns 3 to 8.

Table 4: Expert ratings of questions from question paper 4 with the mark scheme

Demand level (rating category)		*Expert 1*	*Expert 2*	*Expert 3*	*Expert 4*	*Expert 5*	*Expert 6*
Complexity	5 (High demand)	2 3 4 5 7	2 6	4 5 7			1 2 5 7
	4	1 6		1 2 3 6	1	3 4 5 7	6
	3		1 3 4 5 7		3 5 6 7	1 2 6	3 4
	2				2 4		
	1 (Low demand)						
Resources	5 (High demand)	2 3 4 5 6 7	2 5	5		2 3 4 5 6 7	1 4
	4		3 4 6 7	2 3 4 6 7	1	1	
	3	1	1				3 5 6 7
	2			1	2 3 4 5 6 7		2
	1 (Low demand)						
Abstractness	5 (High demand)	2 4 6	2 4 6			2 4 6	2 4 5 6
	4	1 3 5 7	3	2 3 4 5 6 7	2 3 6 7	1 3 5 7	1 3 7
	3		5 7		1 4 5		
	2			1			
	1 (Low demand)		1				
Task strategy	5 (High demand)	1 2 3 4 5 6 7	6	1 2 3 4 5 6 7		4	1 2 3 4 5 6 7
	4		1 2 4 5		3 4 5 7	2 3 4 5 6 7	
	3		3 7		1 2 6	1	
	2						
	1 (Low demand)						
Response strategy	5 (High demand)	4 7		1 2 3 4 5 6 7		3 4	1 2 3 4 5 6 7
	4	2 3 5 6	5		4 5 7	2 5 6 7	
	3	1	1 4 6 7		2 3 6	1	
	2		2 3		1		
	1 (Low demand)						

Note: Question numbers appear in the cells in columns 3 to 8.

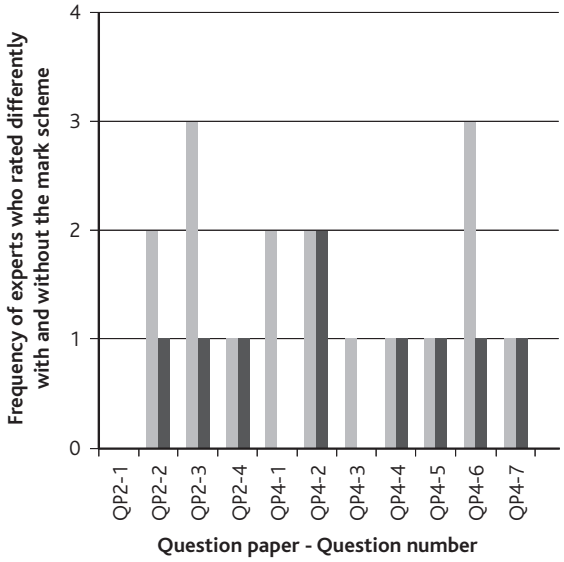


Figure 3: Frequency of experts who rated questions differently with and without the mark scheme for the Complexity dimension

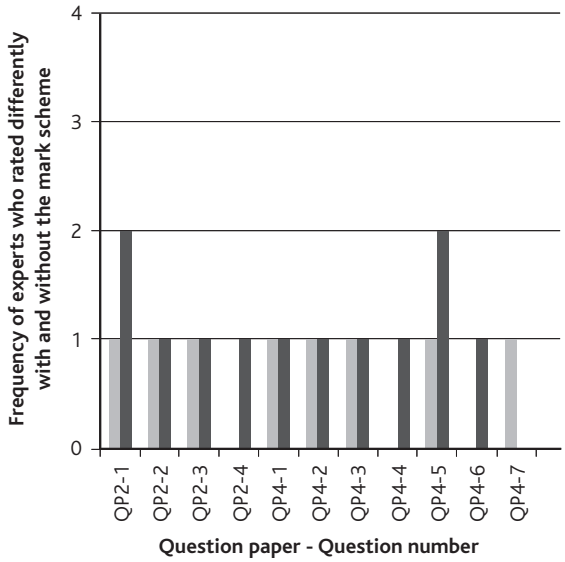


Figure 4: Frequency of experts who rated questions differently with and without the mark scheme for the Resources dimension

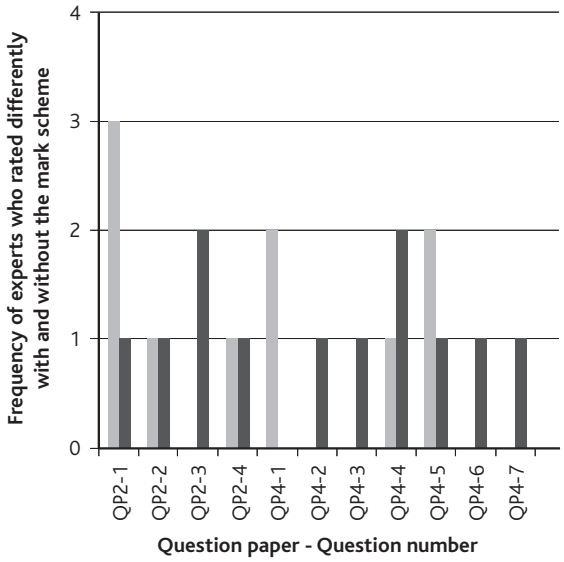


Figure 5: Frequency of experts who rated questions differently with and without the mark scheme for the Abstractness dimension

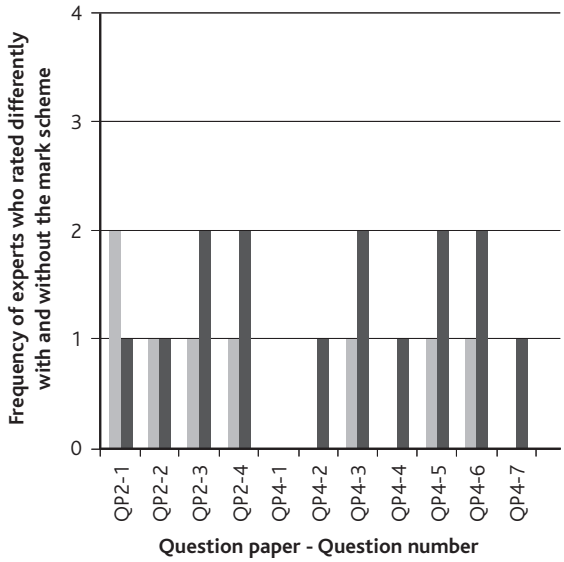


Figure 6: Frequency of experts who rated questions differently with and without the mark scheme for the Task strategy dimension

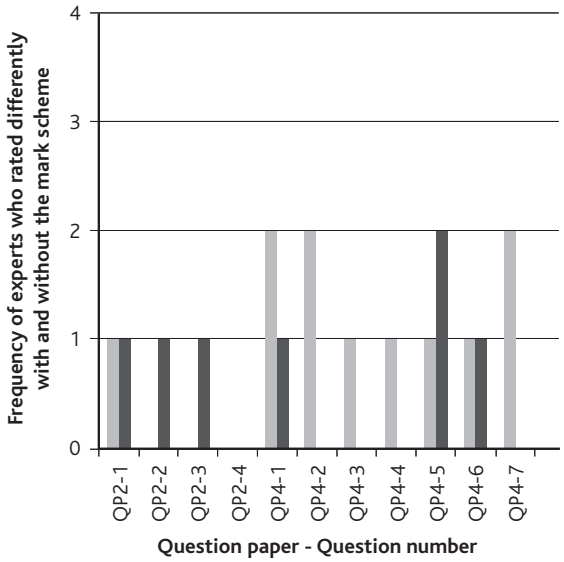
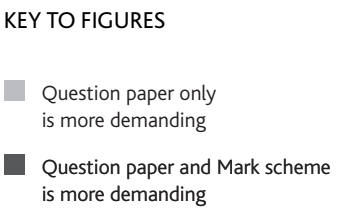


Figure 7: Frequency of experts who rated questions differently with and without the mark scheme for the Response strategy dimension



Reflections on using CRAS with essay and data response questions

The CRAS framework allows the construction of extensive datasets of ratings but tightly circumscribes the methods available for analysing those datasets. Much of the difficulty in analysis stems from the nature of the scales used by raters. Each rater's scale is ordinal so commonly-used descriptive statistics such as means, modes and standard deviations are inapplicable. It is also not possible to compare or combine results from multiple raters on a single dimension without assuming that the raters have a common internalised scale. This assumption is difficult to support and there is no evidence that it holds for this study. Finally, there is no simple method for combining the five dimensions to give an aggregated difficulty score for an item as there is no justification for claiming that, for example, a demand of 3/5 on Resources and a demand of 3/5 on Complexity are equivalent.

One approach which can be pursued is that of making comparisons within a single demand type as rated by a single expert (Johnson and Mehta, 2011). This does allow consideration of the diversity of demand across items, though not in a strictly quantitative manner. It also allows comparison between demands elicited by the question and demands inherent in the mark scheme for a given item. These methods allow somewhat narrow conclusions given the wealth of data from which they are drawn, but the limitations on analysis inherent in the CRAS framework preclude more far-reaching analysis.

The results reported here show low levels of inter-rater agreement in terms of absolute level of demands, but interesting commonalities in terms of ranking of items. For many dimensions, the raters tended to find that all or most items mapped to the same level of demand, but differed strikingly on where that level fell on the 1–5 scale. This suggests a common understanding of how the demand of each item relates to the demand of others around it, but very different anchor points for the internalised scales. Having raters produce a rank order for the items on each dimension, rather than placing them on a scale, could allow finer distinction between items, but would not have revealed the result reported here on the homogeneity of demand across items.

The lack of consistency between raters' internal scales reported here could be related to the lack of an established community of practice. As the raters worked remotely with all information and instructions passing outward from a central hub – the research team – there was no opportunity for raters to negotiate common understandings of the CRAS framework. Though the explicit instructions were common, the tacit, internalised understandings appear significantly divergent. Wolf (1995) argued, in the context of marker reliability, that standards are conveyed through illustrations of students' work within close-knit expert networks, rather than by written criteria. Utilising an analogous process to build a shared understanding of the CRAS scales could move raters towards producing comparable ratings.

Future studies could overcome the lack of consistency between raters' internal scales through alterations to the methods. Particularly effective in avoiding the difficulties inherent in working with CRAS data are approaches based on rankings, such as:

- Collecting data using the Q sort method. Experts would initially work with one CRAS dimension. They would be given each question (or subquestion) in hard copy on a card. They would individually sort the cards, first into two piles, one for 'more demanding' and one for 'less demanding'. After completing this, the questions from the 'more demanding' pile would be sorted into three piles, each corresponding to categories 3, 4 and 5 on the CRAS dimension. Thirdly, the questions in the 'less demanding' pile would be sorted into three piles, each corresponding with categories 1, 2 and 3 on the CRAS dimension. The number of cards in each pile would be restricted to conform to the normal distribution; with most cards in category 3, fewer in categories 2 and 4 and the least in categories 1 and 5. The experts would then rank the questions in each pile from most to least demanding. At all stages the decisions can be reviewed as necessary. Q sort data is generally analysed using cluster analysis to produce statistical summaries of similar Q sorts. The process would be repeated for each CRAS dimension for questions with and without the mark scheme. For more details about the Q sort method see van Exel and de Graaf (2005).
- Collecting data using paired comparisons. Experts would initially work with one CRAS dimension. Each expert would be presented with pairs of questions (or subquestions) and asked to indicate which question in the pair was the most demanding on a CRAS dimension. This would be repeated for all possible pairs of questions. The frequency with which each question was judged to be the most demanding would be used to produce a rank order of questions on a given CRAS dimension. The process would be repeated for each CRAS dimension for questions with and without the mark scheme. Further descriptions of paired comparison methods can be found in Vance and McCall (1934) and Crisp and Novaković (2009a, 2009b).

For both Q sort and paired comparisons each expert would, for each CRAS dimension, produce two rank orders of questions from the most to the least demanding; one with and one without the mark scheme. The frequency of experts who ranked questions differently with and without the mark scheme would be analysed. The Q sort and paired comparisons use rankings rather than ratings, and thereby overcome any leniency or severity in experts' judgements. This avoids issues of differing anchor points and internal scales but it would not standardise the experts' understanding of the CRAS scale.

Finally, the intra-rater approach used to compare the demands of the questions and the demands of the mark schemes provides a robust method for analysing CRAS data. Much of the richness of the original CRAS ratings was lost by calculating the number of experts who rated the question or mark scheme as the more demanding but the method provided useful evidence on the validity questions being addressed. The choice between rigour of analysis and maintaining the richness of the data is common to many decisions made when analysing CRAS data. The use of intra-rater comparisons in the interpretation of CRAS data is also in accordance with Johnson and Mehta (2011).

Reflections on using CRAS with multiple choice questions

The data from the multiple choice papers could not be usefully analysed with the methods used for the other question types and was therefore not presented in the results section. Comparison of the demands inherent in the questions with the demands rewarded by the mark scheme was not possible as the experts could tell what the mark scheme would contain from seeing the question paper. As the number of questions was large (30 per paper) and aggregation of data across questions was not

Table 5: Recommendations for future research in response to experts' comments

Points made by experts	*Suggestions for future research which includes multiple choice questions*
Response strategy is more relevant to open ended questions than closed or multiple choice questions. Experts in a study conducted concurrently with the work reported here experienced similar issues, according to discussion with the authors of Crisp and Hopkin (*in submission*).	Experts rate multiple choice questions on the first four dimensions of CRAS only, thereby omitting Response strategy. Alternatively the equivalent scale developed by Hughes *et al*. (1998) could be used as it makes no distinction between Task strategy and Response strategy.
Students have just two minutes to answer each multiple choice question and the examination setters therefore deliberately avoid complexity and unduly complex reasoning. For instance, calculations requiring more than two steps are avoided. This would lead to low Complexity ratings.	Low complexity on multiple choice papers could be viewed as a legitimate result, reflecting what candidates have to do to respond to the items. It could also be the case that two-step calculations that require "Synthesis or evaluation", "technical comprehension" or making "links between cognitive operations" (Pollitt *et al*., 2007, 186) have high complexity ratings. If the multiple choice papers contain many questions that contain recall alone, however, CRAS might not be the best instrument to apply to them. A pre-rating meeting for experts to agree on an interpretation of CRAS could help with these issues.
Questions are designed to test one topic/problem only. Distractors (incorrect answer options) might relate to a different problem but the student is not expected to solve that problem, just to reject the distractor. This would lead to low Complexity ratings.	The act of choosing between possible responses is not necessarily cognitively undemanding and could include careful evaluation of options – a mark of higher complexity. It could also be that the complexity of items on a given multiple choice paper is generally lower than that on a matched essay paper, which would constitute a useful result.
Questions are designed to include all the necessary information in the question stem. The exception is when a question is intended to test whether the students can determine what is and is not relevant. This would give low Resources ratings.	As above, a tendency towards low ratings does not necessarily constitute an indictment of the method. Some variance in resources ratings would also be expected from the deliberate inclusion of extraneous information in some questions, requiring candidates to select resources.
Questions are designed to test higher order cognitive skills and problem solving skills and this does not seem to be reflected in the rating scale.	This criticism is a direct challenge to the validity of CRAS since if the rating scales do not reflect 'higher order cognitive skills' it is unclear what they do measure. While the CRAS framework has been found to be a useful tool in multiple studies (Hajo, 2008; QCA, 2003, 2006) this suggests the expert's internalisation of the framework was markedly divergent from that which was intended by the researchers. In future work, close attention needs to be paid to establishing with experts the nature of the instrument and good practice in its use.

possible, presentation such as that shown in Table 1 was not practical. The lack of correspondence between the Response strategy dimension and the experts' task also presented problems for analysis. As noted earlier the research evidence prior to the present study found no problems using the Response strategy dimension with multiple choice questions (Shaw and Crisp, 2012).

The following section explores these difficulties through the comments made about the multiple choice task by the raters, and proposes methods for usefully investigating multiple choice questions using CRAS.

Conclusion

In the context of the present study, the CRAS methodology provided validity evidence, though the strength of the evidence provided by the method does not justify the researcher effort required to implement it. The use of expert ratings proved very problematic as consistency of internal scales across examiners is hard to establish, and was not established here. Future work could use rank order approaches to mitigate this difficulty. The aggregation of results from multiple choice items in a manner that produced answers to the research questions was not possible here, and future uses of CRAS might best be restricted to free response items. CRAS could provide a useful tool in validity studies where both the question types and the constructs map to CRAS and either the experts produce rank orders or there is significant commonality among experts in their understanding of the scales and method.

References

Brown, T. (2010). Construct Validity: A unitary concept for Occupational Therapy assessment and measurement. *Hong Kong Journal of Occupational Therapy*, **20**, 1, 30–42.

CIE. (2009). Cambridge International A & AS Level Economics Syllabus code 9708. For examination in June and November 2011.

Crisp, V. & Hopkin, R. (*in submission*). Modelling question difficulty in an A Level Physics examination. *Research Papers in Education*.

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.

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

Ertl, H. & Stasz, C. (2010). Employing an 'employer-led' design? An evaluation of the development of Diplomas. *Journal of Education and Work*, **23**, 4, 301–317.

Hajo, Z. (2008). *Content validity and comparability of the Lebanese national examinations in Chemistry*. Doctor of Education, Leicester, Leicester.

Hughes, S., Pollitt, A. & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A Level exam questions*. Paper presented at the BERA conference, The Queen's University, Belfast.

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, **112**, 527–535.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In: R. W. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications*. 39–64. USA: Information Age Publishing.

Messick, S. (1989). Validity. In: R. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 9, 741–749.

Pollitt, A., Ahmed, A. & Crisp, V. (2007). The demands on examination syllabuses and question papers. In: P. Newton, J. Baird, H. Goldstein & H. Patrick (Eds.), *Techniques for monitoring the comparaility of examination standards*. 166–206. London: Qualifications and Curriculum Authority.

Pollitt, A., Hughes, A., 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.

QCA. (2003). *Report on Comparability between GCE and International Baccalaureate Examinations*. London: QCA.

QCA. (2006). GCSEs and IGCSEs compared: GCSE and IGCSE examinations in 2005 in English, French, mathematics and science (double award). London: QCA.

Quinlan, T., Higgins, D. & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® Scoring Engine *ETS RR-09-01* Retrieved from http://www.ets.org/Media/Research/pdf/RR-09-01.pdf

Shaw, S. D. & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication*, Special Issue **3**.

Shaw, S. D., Crisp, V. & Johnson, N. (2011). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, **19**, 2, 159–176. doi: 10.1080/0969594X.2011.563356

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, **51**, 2, 161–179.

Threlfall, J., Nelson, N. & Walker, A. (2007). *Report to QCA on an investigation of the construct relevance of sources of difficulty in the Key Stage 3 ICT tests* (S. o. E. Assessment and Evaluation Unit, The University of Leeds., Trans.). London.

Tran, H. P., Griffin, P. & Nguyen, C. (2010). Validating the university entrance English test to the Vietnam National University: A conceptual framework and methodology. *Procedia – Social and Behavioral Sciences*, **2**, 2, 1295–1304.

van Exel, J. & de Graaf, G. (2005). Q methodology: A sneak preview. Retrieved from http://qmethodology.net/PDF/Q-methodology%20-%20A%20sneak%20preview.pdf

Vance, T. F. & McCall, L. T. (1934). Children's preferences among play materials as determined by the method of paired comparisons of pictures. *Child development*, **5**, 3, 267–277.

Vogt, D. S., Proctor, S. P., King, D. W., King, L. A. & Vasterling, J. J. (2008). Validation of scales from the Deployment Risk and Resilience Inventory in a sample of Operation Iraqi Freedom Veterans. *Assessment,* **15**, 4, 391–403.

APPENDIX A: Example of a response sheet

Please rate the demands of the activities that the candidates have to do to answer each question where 1 represents 'low demands' and 5 represents 'high demands'. Please do NOT use the mark scheme for this exercise.

Demand type (see detailed information)

Rate from 1 to 5, where 1 represents 'low demands' and 5 represents 'high demands'

		Complexity	*Resources*	*Abstractness*	*Task strategy*	*Response strategy*
Paper 2 questions	Section A Q1					
	Section B Q2					
	Section B Q3					
	Section B Q4					