Evaluating the CRAS Framework: Development and recommendations

Martin Johnson and Sanjana Mehta Research Division

Introduction

This article reviews conceptual issues surrounding comparisons of demand through a critical evaluation of the CRAS (Complexity-Resources-Abstractness and Strategy) framework (Pollitt, Hughes, Ahmed, Fisher-Hoch and Bramley, 1998). The article outlines the origins of the CRAS framework in the scale of cognitive demand (Edwards and Dall'Alba, 1981). The characteristics of the CRAS framework are then outlined, with attention being drawn to the assumptions that underlie these characteristic features. The article culminates in a set of recommendations and guidance that are relevant for potential users of the CRAS framework.

The development of the CRAS framework

The CRAS framework (Pollitt *et al*., 1998) is an adaptation of an earlier scale of cognitive demand (Edwards and Dall'Alba, 1981). The Edwards and Dall'Alba Scale of Cognitive Demand was developed to evaluate lower-secondary level science materials. The primary purpose of the scale was to assess the cognitive demands set within the objectives, the learning tasks, and the evaluation instruments or techniques available to educators and to allow them to evaluate the internal consistency of cognitive demands across these different components. The theoretical foundation of the tool development process was eclectic, drawing on a number of learning theories including Bloom; Bruner; de Bono; and Novak's (1977) interpretation of Piaget.

For Edwards and Dall'Alba the cognitive demand of a task is based on the interaction of four dimensions: complexity; openness; implicitness; and level of abstraction. Moreover, within each of these four dimensions, six levels of demand were defined. The original scale is shown in Figure 1.

Trialling of the scale showed that the tool was useful when teachers reviewed a broad range of educational materials, enabling them to determine the degree of correspondence between their intrinsic cognitive demands. Furthermore, this trialling suggested that the tool was perceived to be advantageous in a number of respects, for instance, its application could lead to:

> *…awareness of features that may otherwise be overlooked; a more accurate and objective reflection of the materials…. and, revelation of the extent to which student performance on the evaluation instruments accurately represents their mastery of what it was intended they learn.*
> (Edwards and Dall'Alba, 1981, p.164)

The Edwards and Dall'Alba scale of cognitive demands was a primary influence on the development of the CRAS scales, which were specifically constructed to examine the effects of structure on demands in GCSE and A level examination items. Pollitt *et al*. defined demands as:

Figure 1: The Scale of Cognitive Demand: Edwards and Dall'Alba 1981

Characteristic Elements of Groups on the Scale				
Dimensions of Cognitive Demand				
Group	*Complexity*	*Openness*	*Implicitness*	*Level of Abstraction*
1	Simple operations	No generation of new ideas	Data are readily available to the senses	Deals with concrete objects or data stored in the memory
2	Require a basic understanding	↕	Data to be operated on are given	Predominantly deals with concrete objects or issues
3	Understanding, application or low level analysis	Limited generation of new ideas	A large part of the data is given but requires generation of the final outcome	↕
4	** ↕	Generation of ideas from a given data base		Corresponds to concrete-abstract transition
5	Analysis and/ or synthesis	Generation of ideas which are original for the student	Data are not available in a readily usable form – must be transformed	Abstract
6	Evaluation	Highly generative	Require a view of the entity in question as part of a more extensive whole	Highly abstract

** The arrows indicate that the characteristic element is intermediate between two more distinct points on the continuum.

> *…requests that examiners make of candidates to perform certain tasks within a question.* (p.6)

According to this definition, demands depend on the question and are the same for all candidates. Pollitt *et al*. articulate the relationship between the concepts of demands and difficulty more directly in their work when compared with Edwards and Dall'Alba. Pollitt *et al*. point out that these judgements of demand are necessarily made in advance of any knowledge about students' performances on such tasks and stand in contrast to their concept of difficulty. For Pollitt *et al*., difficulty is represented by an empirical measure of how successful a group of students are on an item. In contrast to demand, which has 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, pp.105–106).

Pollitt *et al*. (1998) assessed the validity of an examination by comparing the demands set by the examiners in examination items to their overall impression of the responses to those items using the same CRAS scales. This task was undertaken to distinguish between the predicted demands that the examiners had intended when designing the items and the demands that were reflected in student performance on those same items. In this way the presence of the intended demands could be validated through a reflection of actual performance on the item. Without this post-hoc validation the predicted demands would remain untested and lack any ability to support their wider application.

Another adaptation of the original scale led to the inclusion of an additional dimension called 'strategy' into the new framework. The inclusion of this additional scale was supported by an augmentation of the theoretical base of the original Edwards and Dall'Alba scale.

Another contrast between the original Edwards and Dall'Alba scale and CRAS related to the number of levels of demand and the precision of their definition. The original Edwards and Dall'Alba (1981) Scale of Cognitive Demands consisted of a set of dimensions that ranged across six levels of demand. However, in the CRAS scales, the number of levels was reduced to five. In addition, the levels were more loosely defined. In comparison to the inclusion of explicit descriptions for 20 of the 24 dimension levels in the original scale, the new scales contained descriptions for only levels two and four of each dimension; amounting to eight descriptors in total. Hughes *et al*. (1998) suggest that these amendments were necessary to increase the flexibility of the scales, to move it away from its original science-specific context, and to allow judges (examiners) in other subject areas to use their professional *judgement* to make their own subjective comparisons.

These revisions resulted in the development of the CRAS framework which includes the dimensions of: complexity; resources; abstractness; and strategy (Figure 2).

Further revisions of the CRAS scales were then carried out to develop subject-specific scales for judging demands in examination items in History, Geography and Chemistry. Although acknowledging limitations of the CRAS framework in relation to affective and psychomotor demands, these revisions allowed the authors to claim that:

> *The scales can be used to see if the demands of the (i) text books and teaching materials, (ii) national curriculum, (iii) lesson content, (iv) assessment tasks, and (v) marking criteria, are matched.*
> (Hughes *et al*., 1998, p.18)

The features and assumptions underlying the Scale of Cognitive Demands and CRAS

Both sets of cognitive demand scales have a number of similarities and differences in relation to each other. It is important to compare the underlying reasoning which contributes to these similarities and differences.

This article uses two terms to help elaborate this comparison. The superficial and more obvious characteristics of the scales are termed 'features'. The paper goes on to argue that these features are intrinsically linked to sets of 'assumptions' which underlie them. In other words, assumptions are the logical underpinnings of the scales and which help to shape their features.

Figure 2: The CRAS Framework of Demands: Hughes *et al*., 1998

	← *Level* →				
Dimension	*1*	*2*	*3*	*4*	*5*
Complexity The complexity of each component operation or idea and the links between them	←	• Simple operations (i.e. ideas/ steps) • No comprehension, except that required for natural language • No links between operations	← →	• Synthesis or evaluation of operations • Requires technical comprehension • Makes links between operations	→
Resources The use of data and information	←	• All and only the data/information needed is given	← →	• Student must generate the necessary data/ information	→
Abstractness The extent to which the student deals with ideas rather than concrete objects or phenomena	←	• Deals with concrete objects	← →	• Highly abstract	→
Strategy The extent to which the student devises (or selects) and maintains a strategy for tackling and answering the question	←	• Strategy is given • No need to monitor strategy • No selection of information required • No organisation required	← →	• Student needs to devise their own strategy • Student must monitor the application of their strategy • Must select content from a large, complex pool of information • Must organise how to communicate response	→

This section sets out the features and assumptions for both scales. Once key similarities and differences in these features and assumptions are stated there is a brief outline of the claims that are made by each of the respective authors for each set of scales. The shared features (SF), divergent features (DF), shared assumptions (SA), and the divergent assumptions (DA) are described and evaluated in this section.

Shared features (SF)

SF1: The scales are based on an eclectic combination of educational theories

SF2: The scales are used to determine cognitive demand

SF1: The scales are based on an eclectic combination of educational theories

The original scale draws from a range of cognitive and learning theories: because CRAS is based on these original scales, it obviously draws on the same theories. At the same time, the authors of CRAS supplement the original theoretical foundations with more recent work in order to make the scales more applicable to their particular context (examination materials). It is possible that this process of theory building has some problematic elements.

The development of the original Edwards and Dall'Alba scale was based on selection, interpretation and amalgamation of specific theories. The authors justified this interpretative process by arguing that a single theory cannot be all encompassing; they needed to integrate a range of ideas. Although their justification appears reasonable, the exact process of selecting and combining elements from different theories is not entirely clear and raises an important question: can established theories based on their particular central tenets be aggregated in a single tool?

Research related to combining two or more theories into a single theory or conceptual framework is becoming relatively common. In the absence of all encompassing theories, researchers are increasingly identifying the need to construct broader frameworks by combining theories to study complex realities (Radford, 2008; Wedege, 2009; Strauss, 1986). It is suggested that the integration of theories should result in more holistic answers to certain research questions (Tsamir and Tirosh, 2008). Whilst it is accepted that theories originate in specific contexts and provide particular explanations for phenomena, it is also suggested that elements within different theories could complement each other to arrive at a feasible amalgamation (Strauss, 1986). However, it is very important to define the limits of this combination process in order to ensure that the revised theory remains meaningful and relevant.

The process of combining theories needs to be made transparent. More importantly, it also suggests that a researcher will have to carry out an evaluation of each theory that is being considered for integration in a larger framework to determine its goodness-of-fit in that broader framework.

Since CRAS is based on the theoretical framework of the Scale of Cognitive Demands (Edwards and Dall'Alba, 1981) which combined concepts and principles related to learning and cognition from a number of theories, it carries with it some of the ambiguities related to the original development. Whilst Edwards and Dall'Alba (1981) listed the sources from which each of their four demand dimensions were adopted or adapted, the rationale for this selection was not articulated in detail. In the absence of these details the theoretical conceptualisation of CRAS does not lend itself to a critique of the rationale for choosing between the different, and potentially competing theories that were, and that could have been included in the framework. It can only be concluded that combining concepts from different theories is possible, however, the appropriateness of the theoretical framework on which CRAS is established cannot be fully explored.

SF2: The scales are used to determine cognitive demand

Both the Edwards and Dall'Alba and the Pollitt *et al*. scales were created to assess the cognitive demands that are placed on students when engaging with particular tasks. Whilst Edwards and Dall'Alba tie their scale to the scientific learning domain, they suggest that scale application can be used with a diversity of source documents, for example, "The tool is used to determine the cognitive demand levels of the objectives, learning tasks, and evaluation, and to allow a comparison between these" (1981, p.160). On the other hand, Pollitt *et al*. suggest that their adaptation has less learning domain specificity but that it has a tighter focus on specific source documents, for example, for use with assessment items.

Both scales are based to some extent on the taxonomy of learning objectives developed by Bloom (1956). This taxonomy classified learning objectives into three domains, affective, psychomotor, and cognitive. It is notable that both the Edwards and Dall'Alba and CRAS scales focus exclusively on cognitive demands and choose not to engage with either affective, or psychomotor demands.

Divergent features (DF)

DF1: Scale length and level definition
DF2: Attending to constructs

DF1: Scale length and level definition

A contrast between the original Edwards and Dall'Alba scale and CRAS relates to the number of levels of demand and the precision of their definition. The original Edwards and Dall'Alba (1981) Scale of Cognitive Demands consisted of a set of dimensions that ranged across six levels of demand. However, in the revised Hughes *et al*. (1998) framework, the number of levels was reduced to five. In addition, the levels were more loosely defined in the new adaptation. In comparison to the inclusion of explicit descriptions for 20 of the 24 dimension levels in the original scale, the new framework contained descriptions for only levels two and four of each dimension; amounting to eight descriptors in total. Hughes *et al*. (1998) suggest that these amendments were necessary to increase the flexibility of the framework, to move it away from its original science specific context, and to allow judges (examiners) in other subject areas to use their professional judgement to make their own subjective comparisons.

DF2: Attending to constructs

It appears that the relationship of the two demand frameworks to the concept of construct validity differs slightly. In the development work related to the original Edwards and Dall'Alba (1981) scale there is explicit reference to the way that the content, and perhaps by association the constructs, of the science materials were attended to (1981, p.162). In the CRAS development work this link between demands and content/constructs is less clearly articulated.

Whilst the CRAS framework does not explicitly refer to the concept of construct validity in its dimensions it appears that the concept is implicit within the CRAS framework. Construct validity is a concept that test developers and evaluators need to consider. In the CRAS framework the link between demands and content/constructs appears to be more implicit than explicit. Reviewing an item using CRAS involves an analysis of demands in relation to those intended by the item developer. Any discrepancy between the intended and observed demands would indicate that there might be some potential for construct irrelevant variance which would threaten the validity of the item.

Shared assumptions (SA)

SA1: The interaction of multiple demand factors leads to the overall level of demand
SA2: The scales lead to a descriptive, qualitative account of cognitive demand
SA3: The scales enable evaluation of the internal consistency across the different demands
SA4: The scales can be used in conjunction with performance indicators to give insight into the relationship between demands and difficulty

SA1: The interaction of multiple demand factors leads to the 'overall' level of demand

Although both scales include slightly differing sets of dimensions, both conceptualise 'overall' demand in the same way. In line with Edwards and Dall'Alba's (1981) model, Pollitt *et al*. (1998) suggest that the demand dimensions within their CRAS model interact differently with particular features of an examination item. Since overall demand is based on the interdependence of the individual dimensions, changing one aspect of demand in an item might also alter the demands for other dimensions.

SA2: The scales lead to a descriptive, qualitative account of cognitive demand

Both sets of scales facilitate judgements about the demands of tasks which are essentially qualitative or descriptive in nature. Whilst this assumption is somewhat opaque in the work of Edwards and Dall'Alba, e.g. "[application could lead to]…awareness of features that may otherwise be overlooked; a more accurate and objective reflection of the materials" (1981, p.164), this perspective is more transparent in the development of CRAS: "The scales provide a language for examiners to articulate and share discussion, thus building an awareness of those demands…" (1998, p.18). An important implication of this shared assumption is that both scales aim to build a rich description of the demands inherent to a task.

It is important to highlight the point that the accounts generated through these demand frameworks remain at a general level. They do not offer insight into the variability between situations that might have influenced why there could be a difference between what an assessment item intended to do and how a student performed on it. Through triangulation of the projected demands inherent to assessment items, a curriculum, and a mark scheme, the two demand frameworks seek to present a general picture of demands. This analysis remains at the macro-system level and lacks a particular focus on the individual circumstances which might influence student performance. In other words, micro-level variances at teacher and class level within different schools are not a conceptual consideration of the CRAS or the Edwards and Dall'Alba scales. Users of these scales therefore need to bear these limitations in mind if they are interested in gaining such particular insights.

SA3: The scales enable evaluation of internal consistency across the different demands

The Scale of Cognitive Demands is based on the claim that it can be used to identify and compare cognitive demands across related educational components: objectives, learning tasks, and evaluation (Edwards and Dall'Alba, 1981). Similarly, the authors of CRAS claim that analysis of demands using CRAS across several components (text books and teaching materials; national curriculum; lesson content; assessment tasks; marking criteria) can be carried out to determine the degree of match (Hughes *et al.*, 1998). However, the authors of CRAS do not provide any further details on what may be the ideal level of correspondence between demands across these different components.

SA4: The scales can be used in conjunction with performance indicators to give insight into the relationship between demands and difficulty

Implicit to both sets of scales is a relationship between the demands of a task and its level of difficulty. Although this relationship is not considered to be direct, the use of the scales allows insight into the interplay between these two factors. Again, whilst Edwards and Dall'Alba are more vague than Pollitt *et al*. about the concept of difficulty in their work, it can be inferred that they do allude to the relationship between demands and difficulty, for example, "[application of the scales could lead to]…revelation of the extent to which student performance on the evaluation instruments accurately represents their mastery of what it was intended they learn" (Edwards and Dall'Alba, 1981, p.164).

Pollitt *et al*. (1998) conceptualise this relationship in greater depth through discussion of the use of structure in examination items. The term structure can be used to describe item features such as the layout and the number of steps of operations required. Pollitt *et al*. (2007) explain that structure is widely used by examiners to influence the demands of items, and by considering judgements about the demands in such items it is possible to investigate whether these structural features also have effects on any empirical measures of difficulty experienced by students when attempting such items.

Divergent assumptions (DA)

DA1: Item types that the scales can deal with
DA2: The breadth of contexts for scale use
DA3: The capacity of language to describe judgements
DA4: The relative importance of reliability or validity
DA5: The nature of the judgements supported by the scale
DA6: Combining scale judgements
DA7: The role of the scale user

DA1: Item types that the scales can deal with

The Edwards and Dall'Alba scale was designed for use with evaluation items that had objective or multiple choice characteristics. On the other hand, the CRAS framework was developed to be used with a more diverse set of materials. Hughes *et al*. highlight that the CRAS framework was developed to deal with examinations that incorporated a mixture of both structured and essay items (1998, p.18).

DA2: The breadth of contexts for scale use

The Edwards and Dall'Alba (1981) scale was specifically designed to deal with demands in the context of science materials. The CRAS framework was developed to be able to generalise across a variety of subject discipline levels. Hughes *et al*. state that the CRAS development process purposively involved three subjects (History, Chemistry and Geography) so that content coverage spanned "most of the disciplines (mathematical, literary, and physical and social scientific)" (1998, p.18).

DA3: The capacity of language to describe judgements

The Edwards and Dall'Alba scale includes clearly articulated statements along almost all of the points of the rating scales for each cognitive dimension. This implies that the authors believe that language has a capacity to adequately describe qualities of phenomena which can then facilitate judgements to be made against them. This use of rigidly defined criteria contrasts with the approach taken by Pollitt *et al*. for the development of CRAS. The CRAS framework opted to use only two defined scale points for each dimension. This difference in approach reflects Pollitt's concern that trying to use language to encourage absolute judgement making would be useless, since "language, like judgement, is inherently comparative and only approximately quantitative, and the problems of trying to pin down relative meanings with words are well known" (2007, p.189).

DA4: The relative importance of reliability or validity

The Edwards and Dall'Alba tool includes a highly defined cognitive demand scale, which implies that there is a great emphasis placed on how to support the reliable use of the scale. In light of this, a significant portion of the 1981 Edwards and Dall'Alba paper, describing the process of scale development, deals with the issues of establishing inter-rater reliability for use of the scale. Implicit in this process is the sense that attaining high levels of reliable scale use is predicated on good levels of scale user understanding of the scale descriptors. In this way, high reliability is indicative of high validity.

Pollitt *et al*., on the other hand, base their CRAS model on a set of "less stringently defined" cognitive demand scales at levels 2 and 4 of each of the dimensions (cited in: Hughes *et al*., 1998, p.5). The use of fewer descriptors in the CRAS model allows for the inclusion of elements that are relevant to scale users, thereby potentially enhancing the validity of the scale. At the same time, the existence of fewer descriptors heightens the importance of those remaining 'anchor' descriptors since these are needed to align the relative scales of different users into a common framework, since such a scale will always be "implicitly normed relative to the context in which it is being used" (2007, p.189).

Whilst Edwards and Dall'Alba largely avoid the problem of user interpretative variance with regard to the scale descriptors through clear articulation of each descriptor, the CRAS framework is less prescriptive in terms of the standardisation of user interpretation. This lack of prescription is important to highlight since any differences in scale ratings between two judges on CRAS should reflect 'real' differences in the stimuli being compared. An issue arises if inadequate understandings of scale points exist across judges since any variant outcomes might be indicative of differences in the stimuli being judged and/or differences between individual scale users' interpretations of the scales. The potential existence of these two sources of variance require different analytical approaches for scale interpretation than if only one source of variance was being observed (e.g. Cox, 1980, p.408).

DA5: The nature of the judgements supported by the scale

Because the Edwards and Dall'Alba tool comprises sets of clearly articulated statements at different levels of the scale dimensions there might be an inference made that this well-defined scale can support the making of absolute judgements of demand. This contrasts with the loosely defined CRAS scales, which reinforces the concept that individuals' judgements of demand are essentially relative in nature, that is, relative to other defined points on the scale.

DA6: Combining scale judgements

Again, the implied notion that the Edwards and Dall'Alba tool could help to capture 'absolute' judgements of demand has consequences on the potential combination of such judgement outcomes. Since there is an emphasis on the reliability of scaled judgements in the Edwards and Dall'Alba tool there is a suggestion that these judgements possess some mathematical or statistical characteristics. A consequence of this is that individuals' judgements might legitimately be combined in a quantitative fashion to give an overall level of cognitive demand.

This perspective contrasts very clearly with the Pollitt *et al*. view. Reinforcing the point that the dimensions of demand do not possess a quantitative structure Pollitt *et al*. state "despite the use of scales and the collection of numerical ratings the method is still fundamentally a qualitative methodology" (2007, p.192). The practical consequence of this is that "the results of a demand analysis will be to show that different exams make different demands…and it may be possible to say which demands each one requires most of, but it will usually not be possible to aggregate these validly to say that one is more demanding than the other" (2007, p.192).

DA7: The role of the scale user

The structure of relatively well-defined dimension scales in the Edwards and Dall'Alba tool supports its use across other cases, although only in relation to materials from within the context of Science for which it was developed. This contrasts with CRAS which contains loosely defined scales which are intended for use across different subject domains. This difference in structure and intended context means that the role of the scale user is somewhat different. For Edwards and Dall'Alba the well-articulated dimension scales and the clear context expectation constrains the user to ensure that the tool is applied appropriately. In relation to CRAS, the emphasis is on the tool user to establish whether their particular context is suitable for the application of the CRAS scales, and for the consequent modification of those scales.

Conclusion: recommendations and guidance for CRAS use

The identification of the divergent assumptions between the Scale of Cognitive Demands and CRAS is important as these help to explain the different features of the two scales of demands. Through its validation process the expectation of the Edwards and Dall'Alba scale developers is that it should be used as a tool in a very particular way and with little space for the scale user adaption. This contrasts with the CRAS framework since there is more emphasis on the users to adapt the scale for use in their own particular contexts, as long as they adhere to a number of key assumptions. In this way the CRAS scales operate more as a framework than a tool, with the framework resting on two key assumptions: first, that the four CRAS dimensions are used, and secondly, that the ability of judges to make relative judgements is supported by the scales.

This review of the assumptions and features of the CRAS framework leads to a number of recommendations and guidance notes for potential users of the framework. The links to these features and assumptions are referenced in parentheses.

1. The CRAS framework provides a common language to support teachers', examiners' and syllabus developers' conceptualisation and description of demands. The information elicited through the use of the CRAS framework, and the insights gathered, might be particularly important when working in a context where there is a lack of other evidence to draw on, for example, at the beginning of the development of a new assessment (SA2; DA3; DA4).
2. The CRAS framework is essentially qualitative in nature and can be used to profile the nature of cognitive demands for individual users. The rating for each dimension in one stimulus (e.g. an examination item) by an individual user can be used as a basis for comparison across other stimuli by the same individual. This comparison is meaningful because the user is making ratings according to the same underlying reference scale. It is not possible for an individual user to combine the ratings of each dimension to reach an overall 'level of demand'. This overall score is not meaningful as a basis for

comparing different stimuli because the interplay between the different dimensions might have compensatory qualities. By combining the ratings of different dimensions to arrive at a total score the user compromises the qualitative power of the CRAS framework, which aims to demonstrate that each stimulus has different demands and seeks to give the user a language to explicate the nature of those demands (SF2; SA2; DA3; DA6).

3. CRAS recognises the concept that comparisons are based on relative rather than absolute judgements. Moreover, reflecting the complexities of judgement-making processes, the valid and reliable use of the framework relies on there being unidimensional reference scales for each CRAS dimension. Ensuring that this unidimensionality is maintained is perhaps easiest when there is a single scale user, the assumption being that the user will assign meanings to the scale points in a consistent way when rating different stimuli. Whilst the use of a single rater might maximise the reliability of scale application, it might not satisfy the condition for the scales to generate generalisable outcomes. Where multiple judges are involved in making these judgements there needs to be adequate standardisation so that judges' scale use is underpinned by common understandings of anchor criteria. These ratings might then be collectively analysed or subject to numerical treatment, but these treatments need to be meaningfully related to the nature of the data (DA4; DA6; SA2).
4. CRAS may be used in conjunction with other measures (e.g. facility values) to assess the level of difficulty. CRAS can give an insight into the demands that might relate to final difficulty outcomes, but this relationship remains tentative.

   This uncertainty remains for a number of reasons:

   - It is not necessarily the case that there is a direct 1:1 relationship between the CRAS dimensions of demand and difficulty.
   - Initial estimates of demands might also fail to relate well to actual difficulty measures because the concepts identified in items are not recognised in the connected mark scheme. In such cases the identification of such internal inconsistency would be valuable insight.
   - There might be disagreement between the intended/anticipated demands of an item as perceived by a subject expert and those actually experienced by the test taker. This might be due to a number of reasons: there might be factors unknown to the expert, such as teaching effects, that might have influenced the test taker; there might be misapplication of anchor descriptors in the CRAS exercise; and there might be misjudgement on the part of the expert.

   As a result of some of these factors the outcomes generated through a CRAS analysis will tend to be at the level of offering tentative insight into difficulty outcomes (SA3; SA4; DA2).
5. The CRAS framework relies on the users being able to relate their subject-specialist knowledge to the underlying features of the CRAS dimension scales. A precursor to applying the CRAS framework is the mapping of the dimensions to the area of study. This mapping process not only allows the users to demonstrate that the framework is fit for the context of the study, but it also allows adequate anchors on the dimension scales to be developed. This anchoring process is crucial for the scales to be used correctly. Subject-specialist knowledge level is also a crucial factor as this gives validity to the comparisons being made. If a CRAS user has knowledge that is unevenly balanced across two areas of study it will lead to invalid comparisons being made (DA7).
6. CRAS allows descriptions of cognitive demands to be made across a variety of subjects and qualifications. The potential range of application therefore is quite broad. As stated earlier, the relationship between CRAS and the area of study needs to be mapped. Once this mapping is complete CRAS can help to investigate whether the demands that were intended in an item are actually evident (SF2; SA2; SA3; DA7).
7. The rationale for using the CRAS framework is to investigate whether there is internal consistency between different elements of learning and assessment materials. In order to maintain conceptual clarity it would not be recommended that additional measures of cognitive demands be used in addition to CRAS. If a mapping exercise demonstrates that CRAS needs to be extended to include additional dimensions to deal with a context, this process is preferable to using additional sets of measures or alternative cognitive frameworks. By having a singular framework it is easier to compare measures across different elements to investigate internal consistency (DA7).
8. The original intentions of the CRAS framework were to give insight into the dimensions that contribute to item demand, with comparisons then being possible between different items according to their profile of demands. The CRAS framework is less clear about how these individual item characteristics interact when considered at question paper level, and how demands at an overall level might be conceptualised. What appears clear is that the concept of demands at an overall level would necessitate consideration of all of the items that comprise a question paper, and this would mean that selectively sampling items would be invalid.

   The original CRAS scales were used to rate the demands in single items or in item parts. Shifting away from this use might be considered problematic. In their original work Edwards and Dall'Alba make it clear that the cognitive demand of a task is governed by the interaction of different dimensions of demand:

   The level of cognitive demand of a task is determined by the interaction of all of its dimensions. (Edwards and Dall'Alba, 1981, p.159)

   One problem that flows from this is whether it is meaningful to combine sets of qualitative judgements into a 'CRAS score' for a whole paper. If a holistic profile for whole papers is generated by combining the demand scores for each component item, it is possible that the interplay of these item demands is overlooked. In other words, the interplay of individual demands within a question paper makes it problematic to try to combine all the multiple demands and relative compensations into a meaningful outcome which can be used as a point of comparison. For example, placing more or less demanding items at the beginning of an assessment can have an important impact on overall assessment demand; and this potential source of construct irrelevant variance is not captured by a simple aggregation of item demands to construct a measure of demands at a holistic paper level. Whilst it might be argued that CRAS can be used to compare singular items very well, the use of CRAS for multiple items leads to a superficial overview which gives little insight into how to resolve the multiple relationships between such items.

The CRAS dimensions might be used to give a language that can be used to glean an overall impression of the demands of a question paper, but this comparison will be somewhat superficial. Such an analysis will fail to elicit the particularities of the demands and their interrelationships that the framework was initially developed to capture (DA6).

References

Bloom, B. S. (Ed.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Susan Fauer Company, Inc.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, **17**, 4, 407–422.

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.

Hughes, S., Pollitt, A., & Ahmed, A. (1998, August). *The development of a tool for gauging the demands of GCSE and A level exam questions*. Paper presented at the British Educational Research Association Annual Conference, Queen's University Belfast.

Novak, J. D. (1977). *A Theory of Education*. London: Cornell University Press.

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

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.

Radford, L. (2008). Connecting theories in mathematics education: challenges and possibilities. *ZDM Mathematics Education*, **40**, 317–327.

Strauss, S. (1986). Three sources of differences between educational and developmental psychology: resolution through educational developmental psychology. *Instructional Science*, **15**, 275–286.

Tsamir, P. & Tirosh, D. (2008). Combining theories in research in mathematics teacher education. *ZDM Mathematics Education*, **40**, 861–872.

Wedege, T. (2009). *Combining and coordinating theoretical perspectives in mathematics education research*. Proceedings of CERME 6, January 28th-February 1st 2009, Lyon France.

Developing a research tool for comparing qualifications

Jackie Greatorex, Sanjana Mehta, Nicky Rushton, Rebecca Hopkin and Hannah Shiell Research Division

Abstract

There are thousands of diverse qualifications in the UK. Comparability studies about qualification standards generally use the following as comparators:

- Quality of candidates' performance
- Demand

For new and vocational qualifications, samples of candidates' performance and assessment tasks (e.g. examination questions) can be small or unrepresentative and thereby inappropriate for research purposes. Consequently, researchers employ other comparators including *specification features*, e.g. depth of knowledge. The article details the process of devising a research instrument to compare the features of cognate units from diverse qualifications and subjects. Such an instrument is atypical but valuable for comparability studies.

As part of a wider project about comparing different types of qualifications Kelly's repertory grid interviews elicited knowledge from twelve experts. They represented three subjects and composite, general, vocational and vocationally related qualifications. A secondary thematic analysis of the data was completed. The result was a series of features:

- Learning
- Knowledge
- Summative assessment task
- Qualification system

Each feature had several sub-features. Both features and sub-features served to categorise the interview data. An instrument was derived from the features and sub-features, as well as the researchers' experience of qualifications. The instrument was refined through consultation with colleagues. The instrument in its final form consisted of a series of items relating to possible features of the different specifications. Respondents to the instrument were required to tick a box to indicate that the item applied to the given specification. See Appendix 1 for the full instrument.

A pilot of the instrument indicated that salient features vary somewhat between units. Therefore, as hoped, the research instrument highlighted the similarities and differences between units. This is the case for units of the same type and different types. However, there are no established conventions about how to analyse data. Therefore the instrument is suitable for use in future comparability studies about features, as long as the analysis of results is agreed from the outset. Future research might compare qualifications with data collected using the instrument.

Introduction

The aim of this article is to report the development of a research instrument. This is part of an ongoing project about methods of comparing specifications in a diverse qualifications system. For more details see Novaković and Greatorex (2011).

The instrument in its final form consisted of a series of items relating to possible features of the different specifications. Respondents to the instrument were required to tick a box to indicate that the item applied to the given specification. See Appendix 1 for the full instrument. The