5

THE DEMANDS OF EXAMINATION SYLLABUSES AND QUESTION PAPERS

Alastair Pollitt, Ayesha Ahmed and Victoria Crisp

Abstract

Aim

Examiners and many varieties of commentator have long talked about how 'demanding' a particular examination is, or seems to be, but there is not a clear understanding of what 'demands' means nor of how it differs from 'difficulty'. In this chapter we describe the main efforts that have tried to elucidate the concept of demands, and aim to establish a common interpretation, so that it may be more useful in future for the description and evaluation of examination standards.

Definition of comparability

No definition of comparability is necessarily assumed. Sometimes it is apparent that researchers operate with a default assumption that two examinations are expected to show the same level in every aspect of demand, but it would be quite reasonable for one of them to, for example, require a deeper treatment of a smaller range of content than the other; comparability then requires these differences in the demands somehow to balance each other out. It is asking a lot of examiners to guarantee this balance, and a less ambitious approach requires only that the differences are made clear to everyone involved.

Comparability methods

Several methods have been used to look at demands, including: asking informally for impressions of the overall level of demand; asking for ratings of specific demands, or aspects of demand; systematic questionnaires addressing a set of standard demands applicable to many examinations; rating on abstract concepts of demands identified from empirical research. Throughout this work there has been a constant research aspect, as no fully satisfactory system has been developed so far. Theoretical input has come from research in the area, and also from, in particular, taxonomies of cognitive processes, and personal construct psychology.

Strengths and weaknesses

Paying attention to the demands contained within examinations broadens the context of comparability studies, adding a third dimension to comparability. Rather than being just a matter of the ability of the students and the difficulty of the questions, a

focus on demands addresses questions about the nature of the construct being assessed: statistical analysis may tell us that two examination grades are equally difficult to achieve, but it cannot tell us if those grades mean the same thing in terms of what the students who get them can do. We are still, however, trying to develop a system to make this kind of comparison secure, and to establish a common set of meanings to the various terms in use to describe demands.

Conclusion

Demands play an important role in examining in that they are the principal means by which examiners try to control the nature of the construct. When they are constructing the papers and the mark schemes in advance of the test, they have an idea of what the students' minds should be expected to do to achieve a particular grade; by manipulating the demands they try to design tasks that are appropriate for this purpose. To the extent that they succeed, appropriate standards are built into the examination in advance. In this chapter we describe three aims for the study of examination demands. We argue that a description of the nature of the demands is worthwhile in itself, that this can provide a basis for comparing different examinations, and that both of these are valuable even if it is not possible to go further and declare that they differ, or do not differ, in overall demand.

1 Introduction – purpose

> Each review aims to find out if:
> the demand of syllabuses and their assessment instruments (for example question papers, mark schemes) has changed over time
> the level of performance required of candidates at key grade boundaries has changed over time.
>
> QCA (2006a)

These two aspects of standards (they are sometimes called examination demand and grade standard) are commonly considered in most modern studies of examination comparability in England. The quotation refers to standards over time, or longitudinal studies of a part of the system, but cross-sectional studies, where two or more contemporary exams are compared, now also normally address both aspects. In this chapter we are concerned with the first of the two – with the meaning of 'demand' and 'demands' and with how comparability studies have tried to assess and evaluate them.

Before exploring how the present systems for judging examination demands have developed it will help if we start by clarifying the different purposes an assessment of demands can serve. Three separate aims can be identified. First, and particularly if the examinations in question have not been studied much, a purely qualitative study may seek a clear description of the various demands each qualification makes on the students who enter for it. This description is worthwhile in its own right, and valuable to the 'users' of the qualifications: teachers – even students – might use it when choosing which exams to enter for, and employers or other selectors might use it to understand what to expect of those who have taken the exams. This might be called Aim 1: *the aim of description*.

Going further, the aim may be to establish whether or not the exams require similar levels of the demands that they share. This aim is central to the public concern with the maintaining of standards over time, as well as to judging the relative appropriateness of different qualifications for given purposes. This aim needs quantification: a suitable set of scale or construct statements needs to be selected and presented as a set of rating scales to appropriate respondents. If the statements are 'simple' they may be presented to teachers and students as well as examiners, but if they are 'distilled' (as in the CRAS – complexity, resources, abstractness and strategy – system that will be discussed below) the necessary exemplification will make it difficult to involve more than just a team of experienced examiners. The resulting data need to be properly analysed. This might be called Aim 2: *the aim of comparison.*

Finally, demands may be assessed as part of a full-scale comparability study, when the aim of the demands part is to make judges aware of differences in demands before they try to compare grade standards in the second part of the comparability study. Since the link between demands and difficulty, or between demands and performance, is far from straightforward it is necessary to ask them to 'use their judgement' in making allowances for differences in demands when they judge the quality of the work in the scripts they see. Whether judges can, in fact, make appropriate allowances for demands is unclear: in a somewhat similar context Good & Cresswell (1988) found systematic bias when examiners were asked to set equivalent performance standards for examination papers that differed only in question difficulty. Nevertheless, this might be called Aim 3: *the aim of compensation.*

In general, the studies described here do not explicitly state which of these aims they followed. It seems that cross-sectional studies usually adopted the most demanding Aim 3, while, as the Qualifications and Curriculum Authority (QCA) quotation above indicates, longitudinal studies usually expect that the levels of demand will not change significantly over time and aim to test this hypothesis – Aim 2 – before looking at the grade standard.

2 Judging demands

2.1 Demands and difficulty

One significant difference between the general concepts of demands and performance standards is the role of judgement. There is no statistical indicator of demands, and no prospect of our developing objective scales for assessing them. Instead we rely on the judgement of experienced professionals. We could ask the judges to look at students' performances on exam papers and let the evidence of how they dealt with the questions inform the judgements of demands, but in practice we usually do not. We choose to separate the concept of demands from that of difficulty as far as is possible, and ask examiners to use their experience of students' performance on other, similar, questions to imagine how demanding a particular paper 'will' be. Thus 'demands' are also distinguished from 'difficulty' in that the former are essentially a concern pre-test while the other is defined and analysed post-test.

That this distinction needs to be made can readily be demonstrated. Consider, for example, two questions from the Third International Mathematics and Science Study (TIMSS):

1 Subtract: $\begin{array}{r}6000\\-2369\end{array}$

A. 4369
B. 3742
C. 3631
D. 3531

2 Write a fraction that is larger than $\frac{2}{7}$

Answer: ____________________

Location of source material: TIMSS (1996a, pp. 105–7). Reproduced with permission from TIMSS Population 2 Item Pool. Copyright (c) 1994 by IEA, The Hague.

On these questions the success rates of Scottish children were 75% and 76% (data from TIMSS, 1996b, p. 58); thus these questions were equally difficult. In England the success rates were 59% and 79%; clearly they were not equally difficult there. Yet the questions were exactly the same in both countries. Whitburn (1999) gives a plausible explanation for the anomalous English success rate in question 1, in terms of differences in the nature and timing of teaching strategies. Thus, the question was the same in both countries, required the same cognitive operations for its solution, and so made the same demands; but because of differences in their classroom experiences up to the date of the test English pupils found it more difficult than Scottish ones. In essence, by ‘difficulty’ we mean an empirical measure of how successful a group of students were on a question; by ‘demands’ we mean the (mostly) cognitive mental processes that a typical student is assumed to have to carry out in order to complete the task set by a question.

The distinction becomes difficult to maintain when the demand of mark schemes is considered. Students do not see mark schemes – indeed until about 20 years ago they were kept ‘strictly confidential’ and not even released to teachers. A student has two kinds of judgement to make with regard to the mark scheme, corresponding to the two meanings of ‘quality’: what kind of things the examiners are looking for and how good the answer must be to get (say) five marks. If the first of these is problematic the nature of the task is unclear, and this burden of comprehension increases the demands on the student. In the second case, however, there is no extra demand on the student, and it is more appropriate to think of the severity of the mark scheme as an aspect of difficulty rather than demand.

2.2 Simple conceptions of demands

The comments above referred to ‘demands’ as usually understood in comparability studies. There are, however, several features of an examination that can be easily identified, described, and sometimes quantified, as demands. The most frequently mentioned are listed below.

- The amount of *time* spent in assessment varies considerably between subjects, though it is difficult to say whether more time increases or decreases the overall demand. If twice as much time is given because twice as much work is required then the effect is to increase the overall demand, but if more time is given for the same amount of work overall demand will decrease.

- The amount of *work* to be done in that time may vary. In this respect, a paper with more questions in a given time will be more demanding than a similar one with fewer, but if the nature of the questions varies it is harder to quantify the demand. Also, obviously, a syllabus with more content will be more demanding (other things being equal) than one with less.

- More specifically, the amount of *reading or writing* to be done in a given time may vary.

- In addition, the level of *reading difficulty* in the questions may vary, although this is now closely controlled in certificate examinations.

Taken together, these demands may make different examinations more or less suitable for different candidates. In a recent study of vocational tests, the reviewer concluded:

> [Test A] slightly favours candidates whose reading standard is not high; [test B] favours candidates who are more comfortable with intense reading and thinking; [test C] favours those who do not like to be rushed. Unless the typical candidate can be described in more detail, and unless the candidature is unusually homogeneous, it is impossible to say that any of these tests is more or less demanding overall than the others.
>
> QCA (2006b)

- Examination papers may vary in the amount of *question choice* they allow candidates, but because of the interaction of choice with students' expectations, preparation and syllabus coverage the influence of choice on overall demand is complex.

- Subjects vary in the demand they make on *long-term memory*, and so do their examinations. Within subjects, test format and question type can also affect this demand: for example, 'open book' and 'data-book' formats will reduce memory demand, as will information given in synoptic statements in a question, while an essay format may reduce this memory demand, by allowing students to avoid something they can't remember accurately.

- Differences in the nature of questions may also mean that *working memory* demands will vary. This is closely linked to the issue of 'complexity' to be discussed in section 3.

- More generally, the nature of the *cognitive processes* required varies between questions and examinations. This too will be discussed further in section 3.

- For example, candidates undergoing assessment experience *stress*, which can seriously reduce the capacity of working memory. A commonly accepted distinction from Spielberger (1972) holds that stress may be affective, caused by anxiety about the test or its results – such *trait anxiety* is a relatively stable personality characteristic on which individuals vary considerably – or cognitive, caused by the high demands of the context – the ability to deal with *state anxiety* like this is a component of 'expertise'. Some examinations may be more predictable than others, which tends to reduce stress, to reward conscientiousness rather than quick thinking, and to favour students who have been 'well prepared' by their teachers.

Many other factors can affect exam performance. It's clear that a list like this, of features that might make an exam more or less demanding for some students, is in principle endless. For a study of comparability a decision must be made, perhaps on grounds of their possible impact on validity, on which features should be included.

2.3 Explorations of examination demands

Before 1992, only a few comparability studies attempted to consider the demands that syllabuses placed on students, and they used a variety of ad hoc methods to identify specific demands. A study of English language O level by Massey (1979) 'attempted to discern variations in the style and emphasis of boards' questions, including comparisons of the sorts of tasks faced by candidates and an attempted evaluation of their *inherent difficulty* or *complexity of demand*' (p. 2). Views on whether a paper was relatively demanding, relatively undemanding or average were collected from judges (examiners and non-examiners) by questionnaire. Judgements were made on aspects of reading and writing demand – *summary, comprehension, essay* – and *overall demand*. The author emphasised that this method cannot inform about grading standards directly, as awarding can adjust for differences in demands and that 'the comments will be laced with inferences concerning the face validity of examinations, seen from the user's viewpoint' (p. 3) but he considered the issue of interest because exam questions can vary in their complexity.

In a study of A level economics, Houston (1981) asked participants to rate the demands made on candidates as *excessive, appropriate* or *insufficient* by considering the educational aims and objectives, the range and depth of topics and the range of skills specified. When pressed to comment on relative demands, the judges 'suggested that the nine boards offer examinations which make different demands but not necessarily greater or lesser ones' (p. 9). Evans & Pierce (1982) compared the demands of A level German prose composition and free composition between syllabuses. Their analysis of demands was unstructured and was based on the comments and analyses made by the assessors before scrutiny sessions. The analysis considered the weighting of composition, time allocation, length of response, essay choice, the nature of the essay titles and prose passages and how marks were awarded by the mark scheme. Leigh *et al*. (1983) investigated A level music and asked what each board demanded in terms of content and skills, concluding that there was a close underlying convergence of demands.

In 1985, Pollitt *et al.* reported on a study of the sources of difficulty in five Scottish Ordinary Grade examinations, which sought generalisable factors that might be useful to examiners writing questions for the new Standard Grade examinations soon to be introduced. They identified three categories:

1. *subject/concept difficulty*, which relates to the intrinsic difficulty of the content being assessed and the form in which it appears

2. *process difficulty*, related to the psychological operations required to complete the task

3. *question (stimulus) difficulty,* which relates to the wording and other aspects of how the task is presented.

Today we would consider the first two of these to be aspects of demand. The third is, in a very general sense, part of the reading demand, but in practice there are so many specific possible sources of difficulty or easiness in question presentation that it is very hard to make generalisations that would identify them as discrete demands (Ahmed & Pollitt, 1999).

McLone & Patrick (1990) aimed to compare the demands of the two routes available in double mathematics (mathematics/further mathematics or pure mathematics/applied mathematics). Using a matrix based on Griffiths & McLone (1979) a number of statements were presented to judges for rating on a scale of 0 to 3; for example, *How far does the question define in detail the procedure which the candidate should adopt?* Whole papers were then analysed in a similar way after relevant statements had been defined in discussions. The report discussed difficulties with interpreting the different statements, consistency of judgements, using the whole range of ratings and applying the rating scales. The fundamental problem was to define what exactly constituted demand in mathematics, by identifying factors affecting demands and specifying how these factors affect demand. Previous literature had identified three dimensions of demand in mathematics examinations:

1. *academic demand* (intrinsic difficulty)

2. *contextual demand* (demand of the totality of the context within which students are assessed)

3. *personal demand* (contribution to demand of factors relating to the personal characteristics and responses of students).

They considered that the ways in which these interact make it hard to apply scales of demand with precision, and recognised that the actual demands will vary for different participants with different degrees of preparation or familiarity with the materials. It was also noted that there was a lack of empirical data on how factors affect demands and that personal demand will interact strongly with other aspects of demand in somewhat unpredictable ways. Several factors were listed that might, in addition to these dimensions, make questions more or less demanding. It is clear, in retrospect, that *demand* was equated to *difficulty* in this study; today we would translate the phrase *factors affecting demand* to *demands affecting difficulty.*

In general, it was implicitly assumed in these studies that 'demands' should be the causes of difficulty, and that judgements of demands should predict empirical measures of difficulty. But the third category from Pollitt *et al.* (1985), and both the third dimension and the additional factors from McLone & Patrick (1990), show that the difficulty of a particular question is influenced more by very specific features of presentation, and that these aspects of difficulty will affect different candidates in quite different and unpredictable ways.

2.4 Systematic judgements of demands

In 1992 a series of comparability studies was carried out to prepare for changes in GCSE mathematics, English and science consequent to the introduction of the National Curriculum, 5–16. The participating judges were not experienced GCSE examiners, and hence did not have a clear concept of the nature of A-grade work and could not be asked to judge whether a script was above, below or on the grade boundary. Consequently, after initial familiarisation with syllabus materials, judgements of demands were made for each syllabus against a number of defined dimensions. This served as preparation for the cross-moderation phase in which judges were asked to sort scripts into rank-order on each factor (content, context, etc.) and then into an overall rank-order.

For mathematics and science the rating 'factors' were based on Pollitt *et al*. (1985) and work by the Inter-Group Research Committee on 'setting effective examination papers in the GCSE'. In science ratings were made for 'content', 'context', 'processes/skills' and 'question difficulty' (plus 'experimental and practical skills' when considering coursework), while in mathematics ratings were for 'context', 'process' and 'mathematics'. Some differences in the demands of syllabuses were identified. The English judges used the new national criteria for English and English literature as factors. In general the demands were found to be similar though there were differences on some factors. The summary report states that the ratings could 'offer nothing conclusive about comparability (a demanding paper may be generously marked, a less demanding one more severely marked)' but states that 'it provided the context in which to rate the work of the samples of candidates from different groups' (Jones, 1993). Note that this study did clearly separate the concept of demand and difficulty, since the mark scheme was not assessed for 'demand' (cf. section 2.1). Methodologically, there were some problems, which will be discussed in section 4, but a first phase of comparing demands was thought to be a useful and successful addition to comparability studies (Adams, 1993) and was recommended for future studies. There was, however, a feeling that the results of the review and the cross-moderation should be better related in further work.

The comparability studies of 1993 GCSE exams in history (Stobart *et al*., 1994) and geography (Ratcliffe, 1994) included a syllabus/paper review stage with examiners being asked to judge syllabus demands against a number of factors based on Pollitt *et* *al.* (1985). The same method was used for comparability studies in 1994 exams in GCSE mathematics, English and science and A level physics (Alton, 1995; Gray, 1995; Phillips & Adams, 1995; Fowles, 1995). In each study, one examiner from each examining board attended an initial meeting to determine the wording of factor statements to be used and prepare additional guidance on each factor. Every examiner in the study was sent copies of the syllabuses, question papers and mark schemes, and a questionnaire of tables to complete with ratings (1–5) on each factor for each 'foreign' syllabus relative to their own (which they should consider as '3' on each factor). They were encouraged to comment on their ratings, especially at the extremes of a scale.

In general, the factors used were:

- 'content' or 'subject/concept difficulty'
- 'skills and processes'
- 'structure and manageability of the question papers' (question difficulty, language, layout, context, etc.) or 'question difficulty'
- 'practical skills' (in relation to fieldwork) or 'using and applying' (in relation to coursework) – (only used where appropriate).

The range of ratings and mean ratings on each factor were used to compare the demands of syllabuses, and often identified certain specifications as more or less demanding in some ways. Quinlan (1995) used a similar methodology in a study of A level mathematics, but using a list of factors based on McLone & Patrick (1990).

A number of problematic issues were raised, and will be discussed in section 4, but the researchers and the judges were generally satisfied with the methodology (e.g. Stobart *et al.*, 1994; Phillips & Adams, 1995). General satisfaction, however, does not mean that the method was valid and there is a risk that judges may have reported satisfaction just because they were able to carry out the task required of them.

A general caution from several of the study authors warned that the different elements of the studies are not cumulative: 'they provide evidence separately of relative severity or leniency but all three straws pointing in the same direction should not be taken as implying a stronger wind' (Stobart *et al.*, 1994). Differences in demands do not necessarily constitute differences in standards, not least because it does not consider the boundary marks. However, if we were to take on the 'straws in the wind' approach advocated in the 1970s and 1980s (Walker *et al.*, 1987) then consistent outcomes pointing in a particular direction might be taken as more convincing evidence that there is a real difference between specifications, even though they cannot be added up to suggest a larger difference.

In 1996/97 modular and non-modular syllabuses in A level biology, English literature and mathematics were compared. Assessments of demands were made for 'content', 'processes', 'question or stimulus difficulty' and 'modular issues', but the factor statements were finalised by the researchers rather than the judges, and the judges wrote qualitative reports under the four headings instead of making quantitative ratings (D'Arcy, 1997). The judgements were sometimes found to differ because of differing interpretations of the dimensions. Jones (1997) reviewed the methods used and reported several problems mostly centred on the risk of bias arising from judges' familiarity with their own syllabuses or the researchers' summarising of their comments. He concluded that, 'whilst reverting to a tight, quantitative approach was thought not to be desirable, it was considered that more directed guidance, with examples relevant to the syllabuses being reviewed, would enhance this aspect of future studies' (p. 9).

In all these studies, the rating of demands seems to have had two principal purposes: to help ensure that the judges were thoroughly familiar with the materials from all the examination syllabuses before they started the performance judgement task, and to ensure that they could then make appropriate adjustments to their judgements of the quality of performances based on an understanding of any differences in the demands made in each exam. Even when the aim was said to be to 'determine whether or not some of the syllabuses, question papers and mark schemes were perceived as more or less demanding than others' (Stobart *et al.*, 1994) the reason for this was to improve the precision of relative judgements of performance.

In the QCA's Standards Reviews in the late 1990s reviewers were asked to compare sets of examination materials in terms of factors such as: *assessment objectives, rationale, syllabus content, options, scheme of assessment, question papers, tiering*, and *coursework*. In these reviews on behalf of the national regulator, unlike the reports of studies carried out by the examining boards, there does seem to be an assumption that the pattern of demands across alternative syllabuses leading to the same qualification should be comparable – or identical – in its own right.

2.5 Overall demand

Given the complexity of the concept, it is not surprising that very few studies have asked for simple direct ratings of examination demand. When Walker *et al*., (1987) compared A level chemistry between examining boards and over time they mainly used a variety of statistical methods to compare performance standards but also included a judgemental element mainly looking at demands. Examiners were asked to compare the overall demands of each question paper in the syllabus they were involved with to that of the previous year, and to compare the performance of the candidates with the previous year using a five-point Likert scale running from 'considerably higher' to 'considerably lower'. Teachers were also asked to compare the demands of each question paper in the same way. Whilst the data provided an extra source of information for cross-checking the numeric data, the authors acknowledged the limited value that was added given that the judgements were not made between boards and that most studies only look at the examinations in a single year.

More often judges were, and are still, asked to rate overall demand after rating various specific demands, presumably by imagining the overall demand as some undefined composite of these components. This approach and some problems with it will be discussed later. For the moment it can be noted that the conclusion has generally been that the examinations studied have been identified as similar rather than different in overall level of demand.

2.6 Personal construct psychology

During the 1990s a new approach was introduced to considerations of demands, based on the work of Kelly (1955). Personal construct psychology has been defined as:

> ... an attempt to understand the way in which each of us experiences the world, to understand our 'behaviour' in terms of what it is designed to signify and to explore how we negotiate our realities with others.
>
> Bannister & Fransella (1971, p. 27)

According to Kelly, the reality for each individual person is the universe as they perceive it; reality is subjective rather than objective. As they go through life, they actively build up a system of constructs for making sense of the world that is constantly undergoing modification as they experience new events or different outcomes for familiar events. The ability to construe implies the ability to predict (not necessarily always correctly) future events, and so, perhaps, to control one's fate.

Each individual has their own repertory of constructs, and Kelly’s repertory grid analysis is a procedure designed to elicit from an individual how they construe the world. This is the key for our purpose: the repertory of constructs tells us what an individual sees in the world, what is salient, and so offers an insight into what they perceive as demanding in assessment.

Depending on the purpose of the analysis, data may be gathered by eliciting participants’ personal constructs or by supplying them with typical constructs to which they are required to respond. The former approach is necessarily used in psychotherapy, where the concern is for the individual client, and often in the early stages of research; while the latter may be used when the individuals are assumed typical of some population, often in later stages of research. Both methods have been used in comparability studies, to explore constructs and to rate examinations against the constructs that have been discovered.

How are constructs elicited? A construct is, says Kelly (1955, pp. 111–112), ‘a way in which some things are alike and yet different from others’. As a simple example, he gives the statement ‘Mary and Alice are gentle; Jane is not’, which would (probably) be interpreted as indicating that gentleness is a construct that the speaker uses to organise experiences of people. ‘The minimum context for a construct is three things’, he points out: here these are Mary, Alice and Jane. Kelly’s main concern was with human personality, and his therapeutic technique involved asking clients to consider the similarities and differences amongst three people who were significant elements in their lives. But it is not always necessary that all three are mentioned explicitly: ‘To say that Mary and Alice are “gentle” and not imply that somewhere in the world there is someone who is "not gentle" is illogical. More than that, it is unpsychological.' (p. 112). Since in our context the 'elements' would be examination components not well known to the judges, comparisons of three would be difficult for them to cope with, and we generally depend on the presence of the implied third member in each construct elicitation statement.

After eliciting a set of constructs that members of a group typically use, these are defined as bi-polar constructs, such as 'gentle – not gentle' or 'complex – simple'. They are often then combined into a *repertory grid* for further research use. This is a two-dimensional layout with the construct statements listed in the rows and a set of 'objects' heading each column. Using a four- or five-point scale, participants are asked to rate each 'object' on each construct, simultaneously comparing all of the objects on each construct and all of the constructs as applied to each 'object'.

These techniques were originally used in psychotherapy as a means of understanding and thus helping combat patients' psychiatric disorders. Since the mid-1970s it has been applied throughout the social sciences. The first uses in assessment research were in the field of English as a foreign language; Lee (n.d., about 1990) compared the constructs used by a group of Hong Kong lecturers in evaluating writing, and Pollitt & Murray (1993) combined construct elicitation with Thurstone's paired comparison methodology (see Chapter 7) to explore the criteria used by untrained judges evaluating videotapes of speaking tests.

2.7 Use of construct elicitation and analysis techniques

Construct elicitation methods are generally used to identify factors that may differentiate the exam requirements of different syllabus specifications. The first applications involved the 1998 and 1999 GCSE examinations (Gray, 2000; Adams & Pinot de Moira, 2000; Fearnley, 2000; Pritchard *et al.*, 2000) and were followed by a series of studies on 2001 and 2002 AS, A2 and GCE exams (Arlett, 2002; 2003; Edwards & Adams, 2002; 2003; Greatorex *et al.*, 2002; 2003; Guthrie, 2003).

The method typically involves an initial meeting with one judge from each participating board. They compare examination materials (specifications, question papers and mark schemes) from pairs of syllabuses, and are asked to write down similarities and differences (usually a minimum of three of each per comparison) in the demands placed on candidates taking these examinations. Gray (2000) describes this as enabling examiners to form their own ideas of what constitutes demand as a first step in deriving constructs to define a scale of demands. From this a shared set of constructs is agreed by discussion amongst the participants. In a plenary session the wording of the construct statements is finalised (usually formulated as questions), including a title and labels for the ends of each scale. The statements (their number has varied from 14 to 34 in different studies) are then compiled into a questionnaire, to which a final question is usually added asking for a rating of the 'overall demand' of the examination. There is sometimes further refinement through feedback from the judges involved.

A sample of the construct statements that have been generated in various studies, and the bi-polar scale definitions used in the questionnaires, is given below:

1. How accessible are the language and syntax used in the examination papers? Inaccessible – Accessible.
2. What is the predominant type of questions offered to candidates? Short answer – Essay.
3. Is the time allowed for candidates to answer the examination papers enough for them to complete what they have to do? Too little – Too much.
4. Are the assessment criteria for each board equally demanding at grade A? More demanding – Less demanding.
5. To what extent are the questions understandable? Clear – Obscure.
6. What is the role of resource materials? As a prompt – For manipulation.
7. How demanding is the specification in terms of depth? Very demanding – Not demanding at all.
8. Assess the effect upon candidates of increased structure within papers. More demanding – Less demanding.
9. How helpful are the mark schemes to examiners in ensuring consistency in marking? Very helpful – Not helpful at all.

In most studies the construct statements were presented individually, as shown in Figure 1 (Edwards & Adams, 2002). The implication of this format for analysis will be discussed later.

Figure 1 Example of presentation of construct statement

	How accessible are the language and syntax used in the examination papers?						
	Inaccessible						*Accessible*
	1	**2**	**3**	**4**	**5**	**6**	**7**
AQA							
CCEA							
EDEXCEL							
OCR							
SQA							
WJEC							

These construct statements reflect accurately the statements that the judges made during the elicitation procedure. But they vary in several ways – how explicitly they refer to demands, how directly they affect demands, and whether they will affect all students in the same way. The second question seems merely descriptive, though there may be an implicit assumption that some question types are more demanding than others; the fourth question, as discussed in section 2.1, refers more to estimating the difficulty than the demands. The effect of 'increased structure' has been shown to change the nature of the demand in a question, increasing some components while decreasing others, but it is not easy to predict the overall 'effect' (Pollitt *et al.*, 1998), and this sounds here like a request for judges to guess at the difficulty rather than the demands. The terms that define the poles are not always consistent, as when 'Very' is used opposite 'Not at all'.

It has always been the custom to send the judges involved in a comparability study a set of 'familiarisation materials' for each examination they will be judging, including the syllabus specification, question papers, mark schemes and sometimes other documents. Now, in addition, they are sent the demands questionnaire to complete before they attend the main study meeting. The constructs are presented one at a time, with a row of boxes for the different exams being compared. The instrument is thus uni-dimensional, in that the rating a judge gives to one exam needs to be considered relative to the ratings given to the other exams on that same construct.

The questionnaire is not a repertory grid, even though the constructs in it may have been elicited using Kelly's clinical interview technique. A repertory grid is two-dimensional, with all of the constructs presented simultaneously on a single page, with no gaps between them, so that the judge's response procedure is holistic, with each rating being determined by comparisons *both* with other exams on the same construct *and* with other construct ratings for the same exam. The proper analytic techniques are therefore univariate nominal ones for each construct, rather than the specific multivariate techniques developed for repertory grid research.

Analysis of the ratings has therefore often involved the use of chi-square tests within each construct, to check for significant differences between boards. Some studies have then gone on to cluster the examinations in terms of the pattern of ratings given, using a variety of methods (e.g. Gray, 2000; Edwards & Adams, 2002; Greatorex *et al.*, 2003). Greatorex *et al.* (2002) also used the number of constructs for which a board was significantly more or less demanding as an indicator of overall differences in demand. Often analyses have also been carried out to check for bias in the ratings.

The studies often found a number of constructs on which there were significant differences in ratings and a smaller number for which there were significant sub-groups of boards that form clusters with similar patterns of demands. However, significant differences have never been found on the final construct statement rating the overall demands, and stable sub-groups have not been identified where the same boards cluster together on different demands. A typical example was the study of 1998 GCSE English exams (Pritchard *et al.*, 2000), which found no significant difference for 24 of the 37 constructs used. For only two of the constructs were the differences considered 'substantial': on the question 'How effectively is crossreferencing (comparison) tested in the written papers?', specifications divided into two sub-groups with mean ratings of 2.7 and 4.6; and on the question 'How explicitly is the required range of writing targeted in the written papers?', specifications fell into three sub-groups with mean ratings ranging from 2.7 to 4.3. There was no significant difference between the boards in the ratings of overall demand. Similar results have occurred in AS chemistry: Greatorex *et al.* (2002) found significant differences between boards in just 7 of the 24 construct questions involving the transparency of mark awarding, concentration on one or more specification areas in questions, depth of subject content, knowledge required in practical work and emphasis on different areas of chemistry. On the basis of the differences on these, the authors suggested that the Assessment and Qualifications Alliance (AQA) and Oxford Cambridge RSA (OCR) exams were a little more demanding than the others, but reminded us that such an inference should be considered with caution.

Using data published in Fearnley (2000), Baird (1999) explored how the ratings of individual construct statements relate to the ratings of overall demand using forward stepwise regression. Six constructs (only four of which would have been expected to have an impact) were found to explain 60% of the variation in overall demand ratings, but the two constructs found by Fearnley (2000) to be rated significantly differently for different syllabuses were not amongst them. It's noticeable that, in general, the constructs for which the level of demands are found to vary do not appear to be the same in different studies, which perhaps means that different demand components are the most important in the different subjects being examined.

In an attempt to make the demand ratings more accurate and so more informative, Edwards & Adams (2003) allowed examiners to revise their original ratings after the cross-moderation exercise if they wished. Six of the eighteen did so for a few constructs, with changes of up to two or three points on the seven-point scale suggesting that their views changed quite considerably after seeing student work. Most of these changes were made with regard to the Scottish Qualifications Authority (SQA) examination, which was less familiar to most of the examiners. The changes affected the analyses for just two construct statements out of the twenty used in this study, indicating clusters amongst the examinations that had not appeared before. At a research seminar in 2003, Greatorex suggested several ways that might improve the syllabus review method including: make scripts available to help raters understand an examination's demands and to help strengthen the link with the judgement of performance standards; interview judges when they are rating the constructs to improve understanding of how they perceive the cognitive demands; reuse existing scales to systematise the method; analyse verbal protocols collected while judges conduct cross-moderation (reported in Jones, 2004). Jones *et al.* (2004) included one script at each of the A and C borderlines in the familiarisation materials, and used construct statements from previous studies. Judges were also invited to describe any differences they saw in demands and asked to relate these to features of the materials where possible.

Some of these proposals cause us concern. In our view, it is not wise to blur the distinction between 'demands' as the generalisable cognitive requirements that question-writing teams intended to be present in the questions, and 'difficulty' as measured empirically after the event. A few scripts are unlikely to provide reliable 'evidence' to show how students were actually affected by the demands (in our research we have always used at least 200 scripts to look at this), and the empirical outcomes from a question are the proper domain of difficulty, measured statistically, not demands.

The QCA's inter-subject comparability studies (reported in general in QCA, forthcoming a, and also in individual reports), also surveyed elicited demands in four sub-categories: Syllabus, Content, Question papers and their associated mark schemes, and Coursework. An Overall demand rating was also asked for. The instrument used (see Appendix B in QCA, 2006a) asks for a rating of every exam being studied on each scale in turn, so focusing the judges on identifying differences between them. Nevertheless, the overall conclusions were that parallel qualifications were usually similar in overall difficulty, even though there might be substantial differences between them on individual aspects of demand. The science report comments:

> It is clear that awarding bodies working with the regulatory bodies can address a number of these issues through specification review, and guidance on question writing and question paper construction. Specification review is likely to be the first step in order to generate new specifications that recognise the above issues and attempt to do something about them.
>
> QCA (forthcoming b)

This conclusion draws attention to the role that demands play before the examination is seen by students. Question writers, and the scrutiny committees that monitor their work, intend to include appropriate levels of the various demands. Even before that, those who write the syllabus specifications, and the regulators who review them, aim to specify appropriate demands into the examination. If we can establish a consistent system for describing the demands of examinations it can only help writers and reviewers in these efforts.

The general report (QCA, forthcoming a) also noted that the 'reviewers are, by definition, subject experts. However, those taking the papers are, to a large degree, novices. It is a commonplace of examination experience that candidates find questions and sometimes whole papers much harder or easier than those setting them had expected.' Following this lead, Wood & Pollitt carried out construct elicitation interviews with A level mathematics students in which they were asked to describe pairs of questions from AS papers similar to the ones they had recently sat. This study confirmed that students can provide coherent data for exploring the demands of the questions they attempt, and showed that there are significant differences between their and the examiners' perceptions of what makes questions demanding (Wood & Pollitt, 2006).

Most of these studies have reported some problems in using techniques based on personal construct theory, and these will be discussed in section 4. Nevertheless, the methodology has generally been thought to be effective and an improvement on earlier methods, offering a more systematic approach to identifying and comparing demands.

3 Scales of cognitive demands

3.1 Hierarchical taxonomies of demands

The previous section dealt with 'demands' in a very general sense, as any and every challenge that students have to face in certificate assessment. In this section we look specifically at the demands that examination questions make on students' cognitive abilities.

Since the introduction of the O level and O grade examinations it has been standard practice to specify the content of papers in terms of cognitive skills or 'assessment objectives' (AOs). These have generally been derived from the taxonomy of cognitive 'objectives' for education of Bloom (1956), except in the cases of languages, art, and so on (Table 1).

Examination syllabuses often simplify this to two or three levels. A current example (from AQA GCSE chemistry 2007/8) is:

> *AO1* Knowledge and understanding of science and how science works
>
> *AO2* Application of skills, knowledge and understanding
>
> *AO3* Practical, enquiry and data-handling skills

with each of these expanded with three or four specific objectives. The balance of these AOs in each examination component is specified and, increasingly, is mandated by the regulator.

In almost every comparability study judges have looked for differences between examinations in terms of this *intended* pattern of cognitive demands. The QCA review of GCSE history, for example (QCA, 2001), found differences between boards in the percentages of marks awarded for 'low-level skills', 'source interpretation' and 'recall', although it concluded that there was 'a reasonable degree of comparability' overall. Perhaps because of a tightening of the regulators' requirements there is usually very little variation between examinations, at least within similar subjects.

There are very few studies, and no significant comparability studies, where judges have been asked to classify individual questions in terms of Bloom's taxonomy: in general it is either assumed that the examinations were constructed to fit their specifications, or the awarding bodies are asked to provide evidence that they were. Igoe (1982) provides one example of questions being classified cognitively, from the question papers and mark schemes. Items in biology were classified as requiring: data-deduction (numerical or non-numerical), recall (simple, associative or experimental) and logical, coherent argument. However, Igoe did not attempt to

Table 1 Taxonomy of educational objectives. Adapted from Bloom (1956).

Competence	Skills demonstrated
Knowledge	• observation and recall of information • knowledge of dates, events, places • knowledge of major ideas • mastery of subject matter
Comprehension	• understanding information • grasp meaning • translate knowledge into new context • interpret facts, compare, contrast • order, group, infer causes • predict consequences
Application	• use information • use methods, concepts, theories in new situations • solve problems using required skills or knowledge
Analysis	• seeing patterns • organisation of parts • recognition of hidden meanings • identification of components
Synthesis	• use old ideas to create new ones • generalise from given facts • relate knowledge from several areas • predict, draw conclusions
Evaluation	• compare and discriminate between ideas • assess value of theories, presentations • make choices based on reasoned argument • verify value of evidence • recognise subjectivity

measure or compare how demanding items were in different tests. Anderson & Krathwohl (2001) revised Bloom's taxonomy to bring together the knowledge and cognitive process dimensions by mapping them against each other in a two-dimensional framework. The terms of the cognitive process dimension were presented as verbs instead of nouns (remember, understand, apply, analyse, evaluate, create) displayed against the knowledge dimension (factual knowledge, conceptual knowledge, procedural knowledge, metacognitive knowledge). The revisions aim to provide a more authentic tool for planning curriculum, delivering teaching and classroom assessment by helping teachers plan focused objectives. As far as we are aware the revised Bloom's taxonomy has not been used in relation to external assessment, but it may be worth considering how it might be used in at least a descriptive comparison.

Pollitt *et al.* (1985), investigating sources of difficulty rather than demands, rejected the notion of a hierarchy in favour of a list of more specific cognitive processes that might provide a basis for predicting difficulty. The list included:

- explaining
- generalising from data
- selection of data relevant to a general theme
- identifying a principle from data
- applying a principle to new data
- forming a strategy
- composing an answer
- cumulative difficulty
- need for logical consistency.

Examples of most of these were found in each of the five subjects studied.

McLone & Patrick (1990) noted that skilled examiners are able to recognise 'demand' and generally to agree in estimating the overall level of demand in questions. However, they were much less good at explaining it; they could not analyse a question to describe the cognitive elements and processes that were the source of that difficulty. This should not be seen as a criticism of the judges, since they were mathematicians not psychologists, but if we are to arrive at a proper explanation of the demands and difficulties of exam questions, and so to achieve control of this most central element of examining, we need to start by bringing together the expertise of both the subject specialist and the psychologist to develop models for how students think while answering exam questions.

3.2 Analytic scales of demands

Edwards & Dall'Alba (1981) developed and implemented a 'Scale of Cognitive Demand' to quantify the demands placed on the cognitive abilities of students by secondary science lessons, materials and evaluation programmes in Australia. The conceptualisation of demand was derived from a range of learning and thinking theories, including Bloom (1956); Taba (1962, 1967); Bruner *et al.* (1966); Gagné (1970); de Bono (1976); Ausubel *et al.* (1978); and the work of Piaget as interpreted by Novak (1977). Six levels of demand were defined within each dimension, by a list of phrases and command words that were typically used in science textbooks and examinations, or that could be used to describe the processes students were required to carry out. There were four sub-scales:

1. *Complexity*: the nature of the sequence of operations that constitutes a task, that is, the nature of each component operations and the links between operations.

2. *Openness*: the degree to which a task relies on the generation of ideas.

3. *Implicitness*: the extent to which the learner is required to go beyond the data available to the senses.

4. *Level of Abstraction*: The extent to which a task deals with ideas rather than concrete objects or phenomena.

The six levels of 'Complexity' were defined as:

1 simple operations

2 require a basic comprehension

3 understanding, application or low-level analysis

4 [blank][1]

5 analysis and/or synthesis

6 synthesis or evaluation

showing a close resemblance to Bloom's scale. The other sub-scales were new. The scale has not been used directly in Britain.

In a research study conducted for the QCA into the relationship between the increased use of 'structure' in questions and the demands of exam questions the Edwards & Dall'Alba sub-scales were revised to be appropriate for subjects other than science and to be more suitable for rating the demands of exam questions (Pollitt *et al.*, 1998; Hughes *et al.*, 1998). Using insights derived from Pollitt *et al*. (1985) and research into sources of question difficulty (e.g. Pollitt & Ahmed, 1999; 2000), a new trial version of the scales was prepared. This was then revised in discussion with examiners who used it in A level and GCSE chemistry, history and geography, and A level mathematics, and further refined it after a Kelly construct elicitation and repertory-grid rating exercise. The grids were analysed by factor analysis, and revised further.

The final instrument contained four (or five) scales: complexity, resources, abstractness and strategy, and is generally referred to as the CRAS scales.

1. *Complexity* concerned the number of elements that need to be kept in mind while answering, and related to each other.

2. *Resources* related to the extent to which candidates are given all and only the information they need to complete a task, or are required either to supply it themselves or extract it from a source that also contains irrelevant information.

3. *Abstractness* was essentially the same as in Edwards & Dall'Alba.

4. *Strategy* was to assess how much the student was required to devise their own strategies for completing the task. Experience soon showed that the fourth scale should sometimes be split into separate scales called *Problem Strategy* and *Response Strategy,* since exams might differ in the balance of the demands they make on devising strategies for solving problems and on planning how to communicate the answer once it has been found.

It also proved better to define levels 2 and 4, rather than to try to define them all, as Edwards & Dall'Alba had done, or to follow the other common practice of defining the extremes. In later versions some of the statements have been modified to encourage judges to make more use of the extreme categories, such as changing 'No' in the glosses for levels 2 to 'Few' or 'Little'. A current version of the five scales of demands is given in Table 2.

Table 2 The CRAS scales of demands

	1	2	3	4	5
Complexity The number of components or operations or ideas and the links between them.		Mostly single ideas and simple steps. Little comprehension, except that required for natural language. Few links between operations.		Synthesis or evaluation is required. Need for technical comprehension. Makes links between cognitive operations.	
Resources The use of data and information.		More or less all and only the data/information needed is given.		Student must generate or select the necessary data/information.	
Abstractness The extent to which the student deals with ideas rather than concrete objects or 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 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.	

The scales have also been reworded for use in different subjects, with further subject-specific definition to interpret each category to suit each of them. In modern foreign languages, for example, it is stressed that 'resources' refers to the amount and kind of language required from the students in relation to the language they are given in the stimulus material, or to the amount of support they are given for the task. It has been suggested that the scale called *resources* might be better labelled *tailoring of resources.*

Because the descriptions used to define the CRAS scales have been distilled from evidence in many subjects and from many studies into a generic form, one particular application to which they lend themselves is studies comparing the demands and grade standards in different subjects. A series of such studies was carried out by the QCA in recent years, comparing geography to history, the three sciences, media studies to English literature and history, and psychology to biology and sociology (QCA, forthcoming a). In some of these studies ratings were made across qualifications at different levels, from GCSE foundation to A2, the scales were reduced to four levels in each qualification and then overlapped to give as many as ten levels overall.

Ratings were made for every question in one examination paper, plus an overall rating, and this was then repeated for every other exam paper. Note that this contrasts with the usual method in the other studies reported here where all of the examinations were rated together on each scale. One assumes that, with this method, the 'overall' rating will be an implicit average of the ratings of every question, but there is no report of how the judges did arrive at it.

As an example of the findings, in the last of the studies the mean overall ratings were as presented in Table 3.

Table 3 Mean overall ratings

	Biology	**Psychology**	**Sociology**
AS units	2.6	2.8	3.1
A2 units	2.9	4.4	4.2

The report commented:

> …it can be seen that there is very little difference between psychology and sociology at either AS or A2. It can also be seen that both were judged as significantly more demanding than biology at A2 and a little more demanding at AS.
>
> QCA (forthcoming a)

A study of grade standards was also carried out, using Thurstone's paired comparison methodology (see Chapter 7), and the report concluded:

> … the analysis suggested that standards in biology and psychology were very well aligned across the grade range in both the AS and A2 examinations. Given that the initial impulse of the work was the suggestion that students were turning away from science to psychology because it was perceived to be the soft option, the study suggests that this perception has little basis in fact, at least in terms of the demand of the examinations and the grading standards set.
>
> QCA (forthcoming a)

In inter-subject comparability studies it is always going to be difficult to find judges who are capable of rating two or more of the subjects. Having found them, in these studies considerable effort was put into training. An initial briefing preceded the rating; in the fourth study some pilot rating of questions was added to help standardise the ratings. In summary, it seems that the raters did feel confident about their part in the process.

4 Problems in assessing cognitive demands

The reports reviewed in sections 2 and 3 frequently record problems with the assessment of demands. Sometimes these are practical difficulties associated with the particular technique used; others are problems with the principle of the method. Most serious are problems with the conceptualisation of demand, demands and difficulty.

4.1 Practical

Several of these studies noted practical problems that face any attempt to collect ratings of demands. First, these judgements take time, and are therefore an expensive element of a comparability study. The time needed obviously depends to some extent on the number of scales used and the number of times each is applied, and important decisions must be made at the design stage of the study. Attempts to capture the whole of 'overall demand' in a few broad statements means that each statement will be a composite of multiple aspects; whenever these do not correlate highly there will be an averaging effect causing ratings to regress towards the middle category (Fowles, 1995), and real differences between exams may be lost. Time problems are further increased if non-examiners participate (Jones, 1993), since they need more time to familiarise themselves with all of the materials and the assessment procedures before they can judge demands. Yet there are good arguments for using groups other than examiners. Teachers, who prepare students for the examination and are not practised in the arts of question writing, may be in a better position to judge how students will be challenged by a particular feature than examiners who recognise it from past papers. Of course the students themselves are even more likely to understand how demands really operate (Wood & Pollitt, 2006).

The 1–5 numerical scales usually used pose some problems. Phillips & Adams (1995) reported that some raters felt them too limiting; given definitions for '1' or '5' they wanted to expand the scale with '-' and '+' sub-divisions, leading to a 15-point scale. Fearnley (2000) reported difficulties with interpreting qualified descriptors at the ends of scales – how 'few' is 'few' to deserve a '1' rather than a '2'? A similar problem with quantifying features was reported in the 1995 studies: asked to compare 'foreign' exam materials to their own 'home' material that defined the category '3', judges wondered how different the sets needed to be to trigger a rating other than '3'.

4.2 Components

Gray (2000) noted that many of the statements formulated from the comments of judges in initial meetings really expressed simple dimensions of descriptive difference that had little or nothing to do with what most people would consider as demands. In one study as few as 6 out of 14 construct statements seemed to relate to demands. This is a natural outcome of the Kelly elicitation procedure: informants are asked to describe ‘similarities and differences’ they see, not ‘similarities and differences in the demands’. It would be a mistake to ask them to consider whether a difference or similarity concerns ‘demands’ before they speak, since the method depends on spontaneous verbalisation of thoughts, but there is nothing to stop researchers selectively culling the constructs elicited to leave just those that relate to demands.

Simple rating of overall demand, even if it showed differences, would not be very informative, and almost all studies seek ratings of components, or demands. Several reports (e.g. McLone & Patrick, 1990; Jones, 1993) note a concern that as soon as the general concept of ‘overall demand’ is analysed into components there is a problem with potential interactions between the components. Judges reported problems in rating specific demands separately where they believed the total demand would be augmented by interaction. A further complication was added when the ‘style’ of two examinations was deemed different: Jones (1993) and Fowles (1995) both reported that judges found it difficult to make comparative quantitative ratings of demands when this happened. Since these studies concerned GCSE English and A level physics respectively, the notion of ‘style’ clearly must be considered very broadly.

In many studies it is reported that judges had trouble understanding what statements meant. A simple demand like ‘Time available per question’ or ‘To what extent are the questions understandable?’ poses no comprehension problems for judges (however difficult it may be for them to judge it), but the meaning of others, most notably the highly distilled scales of CRAS, may be difficult to master. Greatorex *et al.* (2002) suggest that more discussion between judges is needed to promote a shared understanding of statements like 'How stimulating are the materials?', but Fearnley (1999) argues that even this cannot guarantee consistent interpretations.

4.3 Rating scales

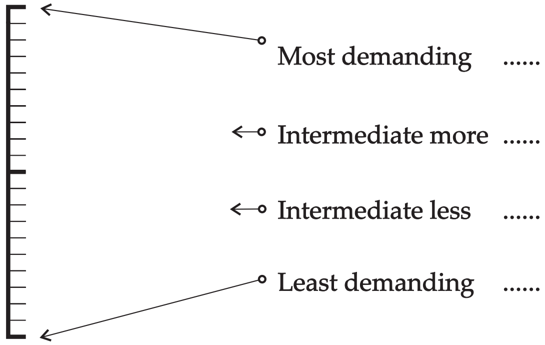
Even if a common meaning could be established for each statement, judges might apply different ‘values’ to the categories within the scale. Every point in the scale needs to be defined quantitatively to avoid this, and this is generally impossible. Language, like judgement, is inherently comparative and only approximately quantitative, and the problems of trying to pin down relative meanings with words are well known. It is not surprising therefore that Adams & Pinot de Moira (2000) question the reliability and validity of some of the data collected.

A further consequence of the comparative nature of judgement (Laming, 2004) is that a 1–5 scale will always be implicitly normed relative to the context in which it is being used. Judges will always tend to place the ‘3’ category, being the middle one, at the centre of what is expected in a particular context. This raises very difficult problems if qualifications at different levels are being compared: GCSE judges and AS judges may both locate ‘3’ as corresponding to the centre of their particular experience, reinterpreting words like ‘usually’, ‘often’ or ‘frequent’ to match their expectation of the average at that level.

A better approach might be devised for controlling the numbers used in the ratings. One study (QCA, forthcoming b) used discussion to partially standardise the rating given by different judges. The same might be achieved more easily by design. For example, consider the form shown in Figure 1: this is designed so that every rater will use the same scale length to represent the four examinations, but will be free to determine the relative sizes not only of the ratings but also of the gaps between them. The ratings will be fully interval, yet will be reasonably well standardised.

Figure 2 A possible standardised rating scale for comparing four exams

Choose the **most** and **least** demanding exam in terms of complexity; then draw lines to place the others on the scale between them:



It is tempting to make the statements narrower in order to reduce comprehension problems, but this can cause other problems, from omitting important aspects of demand to increasing the number of scales and the potential for interactions between them. There is a fundamental dilemma that broad general statements of demand are difficult to understand and rate reliably, but narrow specific ones are not generalisable and their ratings are difficult to evaluate.

The consistency both between and within judges also needs consideration. Methods based on Kelly's technique often result in a wide variance between examiners in the ratings applied to a construct for a particular specification, sometimes covering over half of the full range available and even the full range. This suggests that inter-judge consistency is fairly low. However, it is difficult to assess the general level of inter-judge consistency accurately, as data on ratings have often been reported at the level of the sub-group rather than for each board. Intra-judge consistency has also not been established and it is difficult to guess how consistent an individual judge would be with their ratings if they had made them on a different occasion; further, they might be reasonably consistent in terms of the rank-order in which they place examinations in terms of a particular construct, even if they are inconsistent on a particular rating scale. This suggests further caution in interpreting such methods – as well as the need for some formal investigation of raters' test-retest consistency and construct-specific internal consistency.

4.4 Overall demand from components

The desire to be able to declare one exam to be, overall, more or less demanding than another means that judges are usually asked to make an overall rating. Not surprisingly, they sometimes report problems in doing this (e.g. Edwards & Adams, 2002). Aggregating components without explicit rules is bound to be difficult, but it is unlikely that any acceptable set of quantitative rules could be found.

Arrow's paradox sets requirements for a 'fair' system for aggregating simple preferences into a rank-order and shows that it is impossible to devise a scheme that would always meet these requirements (Arrow, 1951; Vassiloglou & French, 1982). 'Simple preferences' are ordinal measurement, and the impossibility can be avoided if interval data are used. If we ask judges to rate demands on fixed scales we can get interval data, encoding the size of differences between two exams rather than just which is the more 'demanding' on each scale, but it would still be very difficult to obtain agreement on a fair weighting to give each demand. In different examinations, and particularly if they are truly different in style, one would expect different relative importance to be attached to any particular demand[2]. As mentioned before, one solution is just to count how often each exam is deemed more or less demanding than the others (Greatorex *et al.*, 2002).

In the inter-subject studies reported in QCA (forthcoming a) and the related specific reports overall ratings were calculated as the arithmetic averages of the four CRAS scales, which were themselves implicit average ratings given after individual questions had been rated. So long as the 'grand overall average' ratings are treated simply as first indications of potential problems there seems no reason to argue for any more complicated approach than this.

4.5 Construct elicitation technique

There is some concern that the basic presumptions of Kelly's method do not apply in these studies (e.g. Fearnley, 2000). The clinical interview was developed by Kelly in the context of psychotherapy as a method for investigating the mind of a patient. The therapist asks the patient to compare two (or three) people with whom they are thoroughly familiar and about whom they have stable perceptions, such as family members and close friends, and to tell instantly ways in which they are similar or different. In comparability studies judges meet materials for the first time when they are asked to judge them, and it is not obvious that the constructs they express when asked to make comparisons would be the same if they were more familiar with them. However, applied research using Kelly's methods generally involves two phases and it is important to keep them separate.

The first phase is elicitation. In it the comparability researcher is interested not in the mind of the judge but in the constructs elicited from him or her; since the judges are experienced examiners (or experienced teachers or experienced students) they will already have developed the constructs that will allow them to make sense of the examination experience, and it is most probable that they will use these same constructs in the elicitation interview. Of course, if a researcher is still concerned about the unfamiliarity of the materials being used, *since this is just the elicitation phase* it would be acceptable to use only materials familiar to the judges.

The second phase is the rating of the material being studied, using the constructs elicited in the first phase. A wide body of research in psychology (e.g. Fransella & Dalton, 2000; Winter & Viney, 2005), sociology (e.g. Dallos, 1994; Butt & Parton, 2005) and education (e.g. Beard, 1978; Beail, 1985; Pope & Denicolo, 2000) supports the view that the constructs elicited in well-designed interviews do prove valid and useful when used by other judges to rate other similar materials or objects. Most of the studies reported in these use 'repertory grid' techniques, in which the rating data are ordered in two dimensions across both the objects being judged and the constructs being used to judge them. As noted earlier, the comparability studies reported in this chapter generally present constructs singly rather than in a grid, but this in no way invalidates the constructs themselves. Indeed, since the studies do not use repertory grid analytic techniques, they do not depend significantly on Kelly's theory for their validity: his elicitation technique is merely a tool to help set up the scales to be used for judgement.

4.6 Quantification

Houston (1981) and Edwards & Adams (2003) both recognise that the result of a demands analysis will be to show that different exams make different demands. It may be possible to go further and 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 another. It is perhaps easier to see the strength of this argument when the comparison is between different subjects, but it is equally true within one subject.

Arlett (2003) notes that the construct elicitation technique is designed to discover differences (like the Thurstone quantitative technique described in another chapter) and succeeds in doing so. She and others (e.g. Adams & Pinot de Moira, 2000) add that the method provides no way of quantifying or evaluating the significance of the differences it uncovers. This problem gets to the heart of the conceptual confusions that surround 'demands'. Despite the use of scales and the collection of numerical ratings the method is still fundamentally a qualitative methodology, designed to discover and describe differences in the pattern of demands that different qualifications make. Suitable tests can indicate whether or not the differences observed are statistically significant, but they cannot reliably measure their size or educational significance.

Many of the reported problems are a consequence of an assumption that demands and difficulty should be closely linked. McLone & Patrick (1990) saw the fundamental problem as being to identify what constituted demand in mathematics by identifying factors affecting demand and how these factors affect difficulty; one of their categories, 'academic demand', was glossed as 'intrinsic difficulty'. Jones (1993) reports judges' concerns about questions that appeared more demanding than they actually were, the evidence for the latter coming from the mark schemes and marked scripts, and others wanting to see scripts before rating demands because it was difficult to predict how the wording of questions would affect students' work. Reference has already been made in discussing other reports of the usefulness of seeing performance evidence while rating demands. In all of these cases the problem lies in trying to keep separate the two concepts of 'demands' and 'difficulty', and the next section will address this issue directly.

5 Demands and difficulty

5.1 Discussion of the terms

For a student, the outcome of an examination is the grade they achieve, which depends on the score they make and the grade boundaries that are set, and it is generally assumed that the score depends on two factors – the ability of the student and the difficulty of the questions. (This model is discussed further in Chapter 7.) We would like to ensure that the student's grade is determined solely by his or her ability, so that students with more ability always get higher grades, but this will happen only if we can ensure that all of the students respond predictably to the difficulties in the questions. We need, therefore, to understand and control the sources of difficulty in exam questions.

As described earlier, Pollitt *et al.* (1985) identified three kinds of source, which were called *subject/concept difficulty, process difficulty,* and *question (stimulus) difficulty*. We now consider difficulty resulting from the concepts in a subject as aspects of demand; in the CRAS scheme they are rated under 'abstractness' or 'complexity'. Similarly, difficulty arising from the psychological processes the students are asked to carry out is rated as demand in the scales for 'strategy', 'resource', and 'complexity'. For these categories it is fairly simple: more demand quite directly causes more difficulty, and this can be observed as lower scores from students.

The trouble comes with the third kind of source of difficulty. Experience, supported by research (e.g. Ahmed & Pollitt, 1999; 2007; Crisp & Sweiry, 2006), has shown that the differences in difficulty between individual questions depend at least as much on the presence or absence of various features in the stimulus question, as on the amount of difficulty the examiners intended. Some examples from our research in the University of Cambridge Local Examinations Syndicate will illustrate the problems 'questions' cause for examiners.

Example 1

In a GCSE science paper a complicated context was described, which involved a tower for producing fresh water from sea water while also generating electricity. The first part of the question was:

(a) Air rises inside the tower in a convection current. Explain why convection happens.

Many students tried to explain why the tower caused convection – and usually failed. When this was later discussed with the examiners they explained that this was meant to be an easy ‘textbook’ question to get the students started; they did not intend the context to get in the way.

Example 2

It’s commonly assumed that sub-headings will help students structure their answers. This example comes from a GCSE geography paper:

(ii) Describe Gamble Street before urban renewal using the following headings:

open space ______________________________

____

factories ______________________________

______________________________ *(4 marks)*

Performance was disappointing, with students averaging only 33% of the marks. We re-tested the question on a comparable sample without the sub-headings, and – with exactly the same mark scheme – the performance rose to 60%. The students mentioned more of the scoring points listed in the mark scheme when allowed to write more freely.

Example 3

Another GCSE geography question that proved disappointingly difficult, with only 8% success, occurred in a map-reading question:

Using Fig. 1 describe the shape of the valley along this cross-section.

We re-tested this one with the word **shape** printed in bold, and the success rate rose to 37%. The intended task was difficult enough, but the presentation of the question left many students not realising that it was the **shape**, rather than the valley, that they had to describe.

Example 4

The most difficult word in the English language is probably ‘not’. This example is from GCSE mathematics:

Alex, Bernice, Christelle, Divya, Elisa and Fernanda play a game.

They all have an equal chance of winning.

(a) What is the probability that Alex does not win?

In our study sample, 84 students gave the correct answer (5/6) but 93 gave the complementary wrong answer (1/6). Was this because they couldn’t do the maths, or did it result from a reading failure? How can examiners predict the effect of reading failures of this kind?

Example 5

At A level the problems are sometimes quite different; they may be more subtle but equally dramatic. This example is discussed fully in O'Donovan (2005):

Outline ways in which the Conservative Party has rebuilt itself since 1997. (20 marks)

It is hard to blame the examiners for not predicting that some students would challenge the question (but then again, this is A level politics!) and argue that the party had, in fact, failed to rebuild itself; even harder to blame them for not predicting that at least one student would deny that they had even *tried* to rebuild, citing their choices of William Hague and Iain Duncan Smith as proof. It is not easy for examiners to prepare a mark scheme that correctly anticipates the many ways a student may interpret a question, and so to maintain fairness, without sacrificing reliability.

Many more examples could be given, showing how features of language, of layout or of the visual and other resources given in the question, can cause changes in difficulty (usually increases) that are very hard to predict. Examiners deliberately manipulate the demands that contribute to concept and process difficulty and try not to let the presentation of the questions interfere too much with the operation of these demands. Thus the intended demands, mostly in the first two categories of sources of difficulty, represent the trait that the examiners wish to assess. It helps to distinguish the *task* that the examiners want students to tackle from the *question/stimulus* that they use to present it. The sole purpose of the questions is to present tasks that will make the students' minds engage with the intended demands: validity requires that 'the students' minds are doing the things we want them to show us they can do' (Ahmed & Pollitt, 2007), and the question should not prevent that from happening by misdirecting their attention elsewhere.

The many features that can affect the difficulty of the stimulus question could be considered as part of a broad concept of 'reading difficulty', but this is not helpful when it confuses intended and unintended sources of difficulty. We can find no evidence that judges in comparability studies have noted presentation effects like those in the examples above and allowed for them in rating the reading demand of the papers, and it is not reasonable to expect them to do so when the question writers and scrutineers have failed to do so. It is better if the ratings of demands remain as ratings of *intended* demands, where we include in 'intended' any aspect of demand that the question writers could reasonably be expected to have been aware of. The consequence, however, is that ratings of demands will never accurately predict the empirical measures of difficulty derived from students' marks.

Researchers in America have achieved some success in predicting question difficulty from features of questions that might be considered to be demands, but they have been applied only to very limited test-item types, usually testing some aspect of intelligence and set in multiple-choice format (Bejar *et al.*, 1991; Embretson, 1999). Only rarely did they involve language, as in Stenner *et al.*'s (1983) study of vocabulary and spelling.

5.2 Definition of the terms

To summarise, we consider demands to be separable, but not wholly discrete skills or skill sets that are presumed to determine the relative difficulty of examination tasks and are intentionally included in examinations. Examiners use these concepts of demands, fairly deliberately, to control both the nature of the construct that the examination measures and the difficulty of the tasks they use to measure it. Overlaid on the demands, however, is the stimulus question, the layout of words and diagrams (usually) that present the task to the students and that may significantly alter the intended difficulty of the task. Over a whole examination paper, where the same general demands are set at similar levels repeatedly it is likely that the effects of presentation will tend to cancel out, leading to reasonable overall success in controlling the difficulty of examination by focusing on demands. Conceptually, examiners are comfortable talking about the demands in their questions; empirically, because of the powerful influence of the question presentation, it is much harder to confirm their influence on difficulty.

Difficulty, on the other hand, is a statistical measure that indicates how likely it is for any given student to score marks, estimated by considering the scores of actual students in an examination. The difficulty measure is therefore a property of a question or test that is defined for a particular group of students (note how 'difficulty' was reported for different student samples in the TIMSS examples described earlier), and it makes sense to talk of the difficulty of a question and of the difficulty of an examination. The term is also often used loosely to talk of the difficulty of a question for a particular student, but this should be understood as a kind of prediction of the outcome that student (given their ability) can expect on that question (given its difficulty) – again see further discussion of this in Chapter 7.

Amongst the principal differences between demands and difficulty are that the former is judged by experienced participants while the latter is calculated from performance data by statisticians. It is quite important that ratings of demands, conceived in this way, should not be contaminated by performance data since these relate to actual rather than intended outcomes. Allowing judges to revise their ratings after seeing scripts, as in Edwards & Adams (2003), will reduce the value of the ratings as indicators of intended demand, and change them into rather unreliable indices of perceived difficulty. While it may be worth asking judges how difficult they think a paper is, this is not the same as asking how demanding it is; the former is a prediction, the latter a judgement. It is different in an award meeting where the grade boundaries are set. There it is essential that examiners combine their perceptions of difficulty from reading completed scripts with the judgements of demands they made earlier in order to select appropriate scores to act as grade boundaries that maintain the established examination standard.

6 Ways forward

One purpose of this book is to guide future debates about issues relating to examination standards, and a good starting point would be to try to make the use of terms more consistent. The last section developed a definition of *demands*, and the technical definition of *difficulty* presented is well established. It is probably also worth trying to define the noun *demand*: in this chapter it has been used in two senses. We prefer to use it to specify one of the component demands in an examination. But it is also commonly used in a global sense, as the aggregate of all the demands of the whole assessment process; we prefer to describe this as the overall demand. Since demands cannot be quantified (or at least not on a common scale) this aggregation is necessarily subjective, and will be specific to an imagined student group. In these two ways, an examination's *overall demand* is a separate concept from its difficulty. In particular, outsiders will be quite unable to judge accurately the standard of examinations merely by looking at the questions. Experienced examiners who are familiar with the kind of students involved find it hard enough to estimate the overall demand, and without access to the mark schemes they cannot predict the overall difficulty; without also knowing the grade boundaries they have no way of judging the examination standard.

Performance is what candidates actually produce in an examination: it is usually their set of written responses to the tasks set, but it may be some other visible or audible product, and may sometimes be unrecorded; the role of examiners (markers) is to quantify the quality of this performance somehow. The relations between *demands, difficulty* and *performance* are complex. More difficult tasks will usually lead to poorer performances; more demanding tasks may have a similar effect but equally may lead to better performances by prompting students to respond in a more complex or effective way. For similar reasons, changing the nature of demands in a task may raise or lower the measure of difficulty; in addition, since each demand challenges individual students to different degrees, changing the demands may improve some students' performances while worsening others'.

An analogy from sport may help to clarify the relationships between *demand*, *demands* and *difficulty*. In any athletics race it could be said that the *demand* against which candidates are being assessed is 'fast running'. A little thought about the various races makes it clear, however, that there is more to running than this: to compare 100-metre and 5000-metre races we need to consider (at least) two *demands* – 'sprint running' and 'endurance running'. Clearly, some runners are better at one of these than the other. In the longer races we might want to consider 'strategy' as another demand, and for the steeplechase races we might need to add 'jumping or hurdling' demands. A comparison of the marathon and the 400-metre hurdles events would involve judging or comparing them against these various *demands* and would show that one is *more demanding* on some demands and *less demanding* on others. As an extension of this analysis, we could consider the *overall demand* of the event to include some assessment of the amount of training, discipline and sacrifice that successful contestants must accept; but this clearly leads into a very subjective realm of judgement. Such considerations are, however, important in deciding who is likely to achieve most in different events.

The *difficulty* of each race could be measured, as some function of the time, or speed, of winners or of average contestants. This would show that, for example, the 200-metre race was more difficult than the 100-metre; where appropriate, as in qualification for Olympic competition, officials would set suitable 'pass marks' for each event to compensate for the differences in difficulty, using empirical data to determine what is appropriate in each case. The *performance* of each contestant is measured as a time (in other athletics events as a distance). In the decathlon competition empirical data are used to establish rules for rescaling performances for aggregation into a composite total, again compensating, normatively, for differences in *difficulty*. None of these measurements or manipulations, however, affect the *demands*.

There are many other words used, more or less loosely, in discussing examinations, such as the adjectives *demanding* and *difficult*, and other pairs like *challenge/challenging* and *toughness/tough* but it is probably overambitious to try to prescribe how they should be used.

7 Conclusion

This review began by considering the purposes that a study of examination demands might serve, and it is worth revisiting the three aims mentioned in the light of the discussions in the chapter. How well have we achieved Aim 1: *the im of description*? Following a suitable elicitation process, a series of rating scales can be presented to appropriate judges to obtain a description of the intended demands in the exam in as much detail as is desired. This has been done quite successfully in many of the studies reviewed. To improve this, consider the value of such a description: its principal use would be in communicating the nature of the qualification amongst all of the people involved in it – examiners, teachers, students, regulators, employers and selectors.

Can we achieve Aim 2: *the aim of comparison*? The descriptions developed for Aim 1 may be written in terms that are quite specific to all of the participants in that qualification, and which may therefore be misunderstood by others not so closely involved. To meet Aim 2 we need more commonality across the descriptions of different qualifications than there has been so far across the comparability studies. This suggests that it would be worth asking all comparability studies to contribute to a common collection of construct statements, and that some suitable body should undertake to develop from them a standard set of demand scales that can be used in future studies. Eventually this set may be complete enough that there will be no need to carry out an elicitation phase in every study, and it will suffice to select from the construct bank all of the scales that might be important in each new case; as we said earlier the set of demands that are intended to operate in an exam constitute an operational definition of the trait the exam is intended to measure. It should then be easy to make three kinds of comparison. Comparison between different subjects will show how each subject differs in its conceptualisation of achievement, and allow consideration of whether these differences are valuable or problematic. Comparison between examinations that appear to offer the same qualification will help regulators and others to judge whether each qualification is indeed fit for the purposes to which its results are put. Finally, if the description of the demands intended in a given exam are published as part of its specification, a requirement that seems very reasonable when they constitute a definition of what it intends to measure, then comparison between the ratings given to the various demands and the specification will be a form of content validation.

When the GCSE system was being planned in England, several groups tried to 'develop a performance matrix which indicates clearly the attributes that examinations will be seeking to assess and how the levels of achievement will be decided' (Bevan, 1988, p. 1). This is close to an explicit specification for the levels of demand deemed appropriate in each of these examinations, and would provide the basis for holding examinations accountable on demands in just the way that they already are on content.

Finally, how realistic is Aim 3: *the aim of compensation*? There is a problem with this aim, which may originate in what is expected of demands. If, as has been argued here, we should think of demands as 'intended demands' rather than as the 'sources of difficulty' then we cannot logically expect the ratings of demands to predict difficulty very accurately, because of the serious interference of question presentation effects. But if we try to improve the link between 'demands' and difficulty by letting evidence from performance (i.e. evidence about difficulty) modify the judges' initial perceptions of the intended demands then any improvement we obtain will be spurious, being brought about by the very property we are supposed to be predicting.

Remembering that the evidence about demands is essentially qualitative, even if it is expressed numerically on a series of scales, it is probably best not to try to imagine how much 'compensation' should be due to some students because their examination has been deemed more demanding in certain ways than another. Demand differences can be used to test the plausibility of the conclusions from performance comparisons, but we are far from understanding the relationships well enough at present to use them to predict quantitatively differences in performance.

Endnotes

1 Edwards and Dall'Alba gave verbal definitions only for some of their six levels, leaving some blank as 'intermediate' between those above and below them.

2 The problem with Arrow's paradox is rooted in his 'Binary Independence' condition, which requires that the group's relative ranking of two alternatives should not be affected by the presence or absence of any other alternatives (it is often called the 'Independence from Irrelevant Alternatives' condition). Saari (2001) shows that this requirement 'emasculates' the most basic of Arrow's conditions, that all of the judges should behave rationally, by turning all ranked data into a disconnected set of binary comparisons. Since Thurstone's method of paired comparison constructs its scale from exactly such a set of data (see Chapter

7) it follows that Thurstone's method cannot be trusted to meet Arrow's BI condition – in fact, the parameterisaton procedure explicitly contradicts it. Therefore, either we must reject Thurstone's method or we must conclude, with Saari, that Arrow's BI condition is unacceptable.

References

Adams, R.M. (1993). *GCSE inter-group cross-moderation studies summer 1992 examination: English, mathematics and science. Report on the experimental methodology used in the studies*. Welsh Joint Education Committee on behalf of the Inter-Group Research Committee for the GCSE.

Adams, R.M., & Pinot de Moira, A. (2000). *A comparability study in GCSE French including parts of the Scottish Standard grade examination. A study based on the summer 1999 examination. Review of question paper demand, cross-moderation study and statistical analysis of results*. Organised by Welsh Joint Education Committee and Assessment and Qualifications Alliance on behalf of the Joint Forum for the GCSE and GCE.

Ahmed, A., & Pollitt, A. (1999, May). *Curriculum demands and question difficulty.* Paper presented at the International Association for Educational Assessment Annual Conference, Bled, Slovenia.

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualised questions: An experimental investigation of focus. *Assessment in Education*, *14*, 201-232.

Alton, A. (1995). *A comparability study in GCSE science. A study based on the summer 1994 examinations*. Southern Examining Group on behalf of the Inter-Group Research Committee for the GCSE.

Anderson, L.W., & Krathwohl, D.R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: Revision of Bloom's taxonomy of educational objectives: Complete edition*. New York: Longman.

Arlett, S. (2002). *A comparability study in VCE health and social care, units 1, 2 and 5. A study based on the summer 2001 examination*. Organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). *A comparability study in VCE health and social care, units 3, 4 and 6. A study based on the summer 2002 examination.* Organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Arrow, K.J. (1951). *Social choice and individual values.* New York: Wiley & Sons.

Ausubel, D.P., Novak, J.D., & Hanesien, H. (1978). *Educational psychology: A cognitive view* (2nd ed.). New York: Holt, Rinehart and Winston.

Baird, J. (1999, November). *Regression analysis of review outcomes*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Bannister, D., & Fransella, F. (1971). *Inquiring man: The psychology of personal constructs.* London: Croom Helm.

Beail, N. (1985). *Repertory grid technique and personal constructs: Applications in clinical and educational settings.* London: Croom Helm.

Beard, R. (1978). Teachers' and pupils' construing of reading. In F. Fransella (Ed.), *Personal construct psychology. Proceedings of the 1977 international conference on personal construct psychology, University of Oxford.* (pp. 69–74). London: Academic Press.

Bejar, I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

Bevan, D. (1988). *Report of the MEG performance matrices (science) working group: Interim report*. Cambridge: Midland Examining Group.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners. Handbook I: Cognitive domain*. New York: Longmans, Green.

Bruner, J., Olver, R.R., Greenfield, P.M., Rigney Hornsby, J., Kenny, H.J., Maccoby, M., *et al*. (1966). *Studies in cognitive growth*. New York: John Wiley.

Butt, T.W., & Parton, N. (2005). Constructivist social work and personal construct theory. *British Journal of Social Work*, *35*(6), 793–806.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, *48*, 139–154.

Dallos, R. (1994). *Family belief systems, therapy and change*. Milton Keynes: Open University Press.

D'Arcy, J. (Ed.). (1997). *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

de Bono, E. (1976). *Teaching Thinking*. London: Maurice Temple Smith.

Edwards, E., & Adams, R. (2002). *A comparability study in GCE AS geography including parts of the Scottish Higher grade examination. A study based on the summer 2001 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint

Council for General Qualifications.

Edwards, E., & Adams, R. (2003). *A comparability study in GCE Advanced level geography including the Scottish Advanced Higher grade examination. A study based on the summer 2002 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

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.

Embretson, S. (1999). Cognitive psychology applied to testing. In F.T. Durso (Ed.), *Handbook of applied cognition* (pp. 629-660). New York: Wiley.

Evans, B.F., & Pierce, G.E. (1982). *Report of a GCE inter-board study in German at the Advanced level 1980*. Cardiff: Welsh Joint Education Committee.

Fearnley, A. (1999, November). *Kelly's repertory grid and analysis of ratings*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Fearnley, A. (2000). *A comparability study in GCSE mathematics. A study based on the summer 1998 examination*. Organised by the Assessment and Qualifications Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.

Fowles, D.E. (1995). *A comparability study in Advanced level physics. A study based on the summer 1994 and 1990 examinations*. Northern Examinations and Assessment Board on behalf of the Standing Research Advisory Committee of the GCE boards.

Fransella, F., & Dalton, P. (2000). *Personal construct counselling in action* (2nd ed.). London: Sage Publications.

Gagné, R.M. (1970). *Conditions of learning*. New York: Holt, Rinehart & Winston.

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, *14*, 263–281

Gray, E. (1995). *A comparability study in GCSE English. A study based on the summer 1994 examinations*. Midland Examining Group on behalf of the Inter-Group Research Committee for the GCSE.

Gray, E. (2000). *A comparability study in GCSE science 1998. A study based on the summer 1998 examination*. Organised by Oxford Cambridge and RSA Examinations (Midland Examining Group) on behalf of the Joint Forum for the GCSE and GCE.

Greatorex, J., Elliott, G., & Bell, J.F. (2002). *A comparability study in GCE AS chemistry: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2001 examination*. Organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Greatorex, J., Hamnett, L., & Bell, J.F. (2003). *A comparability study in GCE A level chemistry including the Scottish Advanced Higher grade. 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 Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Griffiths, H.B., & McLone, R.R. (1979). *Qualities cultivated in mathematics degree examinations*. Southampton: University of Southampton.

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 examination*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Houston, J.G. (1981). *Report of the inter-board cross-moderation study in economics at Advanced level*. Aldershot: Associated Examining Board.

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, The Queen's University of Belfast.

Igoe, R.A. (1982). *A comparative analysis of the marks awarded in the 1979 Advanced level biology examination of the Joint Matriculation Board and a study of the factors affecting candidates' responses to different types of question.* Unpublished MSc dissertation, University of Warwick.

Jones, B.E. (1993). *GCSE inter-group cross-moderation studies 1992. Summary report on studies undertaken on the summer 1992 examinations in English, mathematics and science*. Inter-Group Research Committee for the GCSE.

Jones, B.E. (Ed.). (1997). *A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Jones, B.E. (2004). *Report of the JCGQ research seminar on issues related to comparability of standards, 3 December 2003*. Internal Research Paper RC/264. Manchester: Assessment and Qualifications Alliance.

Jones, B., Meadows, M., & Al-Bayatti, M. (2004). *Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003*. Assessment and Qualifications Alliance.

Kelly, G.A. (1955). *The psychology of personal constructs* (Vols. I and II). New York: Norton.

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

Lee, Y.P. (n.d.). *Where markers of essays agree and where they don't – An application of repertory grid analysis*. Unpublished, University of Hong Kong.

Leigh, E.M., Ayling, M., Reeve, R.G., Kingdon, M.J., & Ibbotson, P.M. (1983). *A survey of GCE music A level syllabuses and examinations*. London: University of London, University Entrance and School Examinations Council.

Massey, A.J. (1979). *Comparing standards in English language: A report of the cross-moderation study based on the 1978 Ordinary level examinations of the nine GCE boards*. Bristol: Southern Universities' Joint Board and Test Development and Research Unit.

McLone, R.R., & Patrick, H. (1990). *Standards in Advanced level mathematics. Report of study 1: A study of the demands made by the two approaches to 'double mathematics'*. An investigation conducted by the Standing Research Advisory Committee of the GCE Examining Boards. Cambridge: University of Cambridge Local Examinations Syndicate.

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

O'Donovan, N. (2005). There are no wrong answers: An investigation into the assessment of candidates' responses to essay-based examinations. *Oxford Review of Education*, *31*, 395–422.

Phillips, E., & Adams, R. (1995). *A comparability study in GCSE mathematics. A study based on the summer 1994 examinations.* Organised by the Welsh Joint Education Committee on behalf of the Inter-Group Research Committee for the GCSE.

Pollitt, A., & Ahmed, A. (1999, May). *A new model of the question answering process*. Paper presented at the International Association for Educational Assessment Conference, Bled, Slovenia.

Pollitt, A., & Ahmed, A. (2000, September). *Comprehension failures in educational assessment.* Paper presented at the European Conference on Educational Research, Edinburgh.

Pollitt, A., Entwistle, N.J., Hutchinson, C.J., & de Luca, C. (1985). *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.

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., & Murray, N.L. (1993). What raters really pay attention to. Language Testing Research Colloquium, Cambridge. Reprinted in M. Milanovic & N. Saville (Eds.), (1996), *Studies in language testing 3: Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

Pope, M.L., & Denicolo, P. (2000). *Transformative education: Personal construct approaches to practice and research*. Chichester: Whurr/Wiley.

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 examinations*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Qualifications and Curriculum Authority. (2001). *Five year review of standards: GCSE history*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2006a). *QCA's review of standards: Description of the programme.* London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2006b). *Comparability study of assessment practice: Personal licence holder qualifications.* London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (forthcoming a). *Review of standards between subjects: General report*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (forthcoming b). *Review of standards between subjects: Science report*. London: Qualifications and Curriculum Authority.

Quinlan, M. (1995). *A comparability study in Advanced level mathematics. A study based on the summer 1994 and 1989 examinations*. University of London Examinations and Assessment Council on behalf of the Standing Research Advisory Committee of the GCE Boards.

Ratcliffe, P. (1994). *A comparability study in GCSE geography. A study based on the summer 1993 examinations*. Northern Examinations and Assessment Board on behalf of the Inter-Group Research Committee for the GCSE.

Saari, D.G. (2001). *Decisions and elections: Explaining the unexpected.* Cambridge: Cambridge University Press.

Spielberger, C.D. (1972). *Anxiety: Current trends in theory and research: I*. New York: Academic Press.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*, 305–316.

Stobart, G., Elwood, J., Jani, A., & Quinlan, M. (1994). *A comparability study in GCSE history: A study based on the summer 1993 examinations*. University of London Examinations and Assessment Council on behalf of the Inter-Group Research Committee for the GCSE.

Taba, H. (1962). *Curriculum development: Theory and practice*. New York: Harcourt Brace & World.

Taba, H. (1967). *Teacher's handbook for elementary social studies*. Reading, MA: Addison-Wesley.

Third International Mathematics and Science Study. (1996a). *TIMSS mathematics items: Released set for population 2 (seventh and eighth grades).*

Third International Mathematics and Science Study. (1996b). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

Vassiloglou, M., & French, S. (1982). Arrow's theorem and examination assessment. *British Journal of Mathematical and Statistical Psychology*, *35*, 183–192.

Walker, N.A., Forrest, G.M., & Kingdon, J.M. (1987). *Comparing the boards' examinations: An alternative approach*. Report on the second part [1984–1986] of an exercise conducted by The Standing Research Advisory Committee of the GCE boards involving the examinations in chemistry at Advanced level from 1978 to 1982 and from 1984 to 1986 inclusive.

Whitburn, J. (1999). Why can't the English learn to subtract? In B. Jaworski & D. Phillips (Eds.), *Comparing standards internationally* (pp. 163-182). Oxford: Symposium Books.

Winter, D.A., & Viney, L.L. (2005). *Personal construct psychotherapy: Advances in theory, practice and research*. Chichester: Whurr/Wiley.

Wood, A., & Pollitt, A. (2006, November). *Developing a methodology to enable students to participate in test development*. Paper presented at the Association for Educational Assessment – Europe Annual Conference, Naples, Italy.

COMMENTARY ON CHAPTER 5

Alison Wood

Judgements in formal comparability studies are made by experts: subject specialists, many of whom are senior examiners, with years of experience in question-setting and marking. In their chapter, Pollitt *et al.* refer to a study which took a different approach (Wood & Pollitt, 2006). In this study, A level mathematics (statistics) students made comparative judgements about the overall demand of questions, then went on to identify and describe the factors which they judged to impact on it, that is, they identified the particular demands of the questions.

Pollitt *et al.* acknowledge the general point that students themselves will have a particular perspective on the demands of questions. I suggest that a stronger claim might be made: that students are likely to perceive the demands of particular questions in ways which are inaccessible to experts and, for this reason, we should incorporate their judgements into formal comparability studies. If we want to know what the real demands of questions are, then we need to elicit judgements from those who experience them, that is, the students themselves.

Wood & Pollitt (2006) found that, when describing the demands of questions, although there was much agreement between students and experts, the students identified factors which the experts simply did not recognise. Nathan & Koedinger (2000) reported similar findings from their study of lower-secondary algebra students, so our findings were not unanticipated and the literature on problem-solving suggests that such differences are to be expected. This is because there are important differences between experts and novices in the ways in which they perceive and then engage with problems. Experts and novices represent problems differently (Chi, Feltovich & Glaser, 1981). Novices are far less able to identify and represent problems as being of particular types (Chi, Feltovich & Glaser, 1981; Cummins, 1992; Hinsley, Hayes & Simon, 1978; Mayer, 1982; Riley, Greeno & Heller, 1983; Schoenfeld & Hermann, 1982; Silver, 1981), so are less likely to activate the correct problem-solving schema (Paas, 1992). Experts, on the other hand, activate the appropriate problem-solving schemata very quickly, sometimes as soon as the first phrase of the problem statement is read (Hinsley, Hayes & Simon, 1978). This means that, as soon as an expert begins to read a question, s/he recognises which aspects are relevant, organises those aspects into a coherent model and integrates that model with existing knowledge, in order to solve the problem. Quilici & Mayer (2002) refer to this as structural awareness and it is this structural awareness that makes problem-solving much less demanding, for an expert, than for a novice.

Experts also differ from novices in the ways in which they engage with questions when actually working through the question and generating a response. This is because experts find it easier to recall knowledge during problem-solving, because knowledge in their working memory has strong links to chunks in long-term memory. This facilitates recall, with experts recalling not only more information, but also recalling it in an immediately meaningful way (Larkin, McDermott, Simon & Simon, 1980; Lavigne & Glaser, 2001, in the specific context of statistics). Novices recall information in smaller units and chunk it according to more superficial aspects of the information (Chase & Simon, 1973; Feltovich, 1983). Representing a problem and then engaging with a problem are, therefore, different for experts and novices and so their experience of the demand(s) of the problems will differ. It is for this reason that I am proposing that student judgements be incorporated into comparability studies.

This raises the question of which of the chapters' aims the students might be able to address. Beginning with *the aim of description* – laying bare all of the intended construct-relevant demands which a qualification or an examination paper presents – it seems unlikely that students could describe the *intended* demands of a qualification as a whole, as this ought to be a matter for the curriculum/assessment expert. They did seem, however, to be able to describe the demands of questions, in the sense expressed in Chapter 5: 'the (mostly) cognitive mental processes that a typical student is assumed to have to carry out in order to complete the task set by a question'. They were also able to identify construct-irrelevant sources of difficulty. Wood & Pollitt (2006) found that they could do this consistently, indicating inter-rater reliability and there were some demands/sources of difficulty that were identified only by the students.

Turning to the *aim of comparison* – where judges compare the intended demand profiles between two or more examination papers, highlighting similarities and differences – Wood & Pollitt's (2006) students were able to make comparative judgements about question pairs very easily and give reasons for their comparative judgements, again with evidence of inter-rater reliability. Making judgements at whole-question paper level is a (logical) extension of making judgements about question pairs and a small-scale study, carried out as preliminary work for Wood (2006) suggested that students could compare whole question papers, even between subjects. Again, some of the judgements they made differed from those of experts. If students can provide evidence which cannot be generated by experts, it would seem counter-productive not to include them in comparability studies. At best, with appropriate support, they might well be able to address the aims of *description* and *comparison* in ways which are similar to those of expert judges. At the very least, though, their descriptive and comparative judgements should be made available to experts.

The aim of compensation brings together the notions of construct-relevant demands and construct-irrelevant sources of difficulty. In the formal comparability studies, expert judges make an estimate of the overall demand of an examination paper and then combine that with a further estimate of the difficulty of that particular paper. This judgement is made in the context of the particular sample of students who take that examination. This process gives rise to an estimate of how, on average, those particular students will have experienced that paper. The 'average experienced difficulty' judgement which arises from this process is then used to compare one paper with another.

To address *the aim of compensation*, students would need to be able to make this extremely complex 'average experienced difficulty' judgement. Students would not have the experience to enable them to contextualise their demand judgement, but this judgement, if fed into the deliberations of expert judges, could enable those experts to make their 'average experienced difficulty' judgements more reliably.

Quite clearly, thinking about the use of students in comparability studies is at an early stage and raises many questions. Wood (2006) identifies the range of issues requiring further research, focusing on validity (for example, whether the demands identified by the sample of students really did have an impact on (a) their performance and/or (b) the performance of the whole cohort) and generalisability (for example, whether students from a wider range of ages/levels of ability can make demand judgements). She proposes a programme of work to investigate such issues further.

References

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Cummins, D.D. (1992). Role of analogical reasoning in the induction of problem solving categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 1103–1124.

Feltovich, P.J. (1983). Expertise: Recognising and refining knowledge for use. *Professions Education Researcher Notes*, *4*(3), 5–9.

Hinsley, D.A., Hayes, J.R., & Simon, H.A. (1978). From words to equations: Meaning and representation in algebra word problems. In P.A. Carpenter & M.A. Just (Eds.), *Cognitive processes in comprehension*. Hillsdale New Jersey: Erlbaum.

Larkin, J.H., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*, 1335–1342.

Lavigne, N.C. & Glaser, R. (2001). *Assessing student representations of inferential statistics problems*. CSE Technical Report 553. CRESST/Learning Research and Development Centre, University of Pittsburgh.

Mayer, R.E. (1982). The psychology of mathematical problem solving. In F.K. Lester & J. Garofalo (Eds.), *Mathematical problem solving: Issues in research*. Philadelphia: The Franklin Institute Press.

Nathan, M.J., & Koedinger, K.R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction*, *18*(2), 209–237.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*, 429–434.

Quilici, J.L., & Mayer, R.E. (2002). Teaching students to recognise structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*, 325–342.

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Developmenst of children's problem-solving ability in arithmetic. In H.P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). San Diego, CA: Academic Press.

Schoenfeld, A.H., & Hermann, D.J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *8*, 484–494.

Silver, E.A. (1981). Recall of mathematical problem information: Solving related problems. *Journal for Research in Mathematics Education*, *12*, 54–64.

Wood, A. (2006). *What makes GCE A level mathematics questions difficult? The development of the Preferred Alternative Construct Elicitation (PACE) methodology for enabling students to make and give reasons for demand judgements: The findings from a pilot study and an outline programme of work arising from the pilot study*. Unpublished Masters in Research Methods dissertation. University of London Institute of Education.

Wood, A., & Pollitt, A. (2006, November). *Developing a methodology to enable students to participate in test development*. Paper presented at the Association for Educational Assessment – Europe Annual Conference, Naples, Italy.

RESPONSE TO COMMENTARY ON CHAPTER 5

Alastair Pollitt

I fully agree with these comments. Even the best teachers sometimes struggle to see why students find a problem difficult – it is not easy for an expert to 'think like a novice'.

We need not stop with students. There are also good arguments for inviting other groups to take part in certain kinds of comparability study. Suppose, for example, we want to compare the standard of two examinations in Spanish. Who could be better able to judge the communicative quality of students' speech in Spanish as a foreign language than native Spanish speakers, preferably with no knowledge of English and not trained in teaching? If the purpose of language teaching is to enable communication with speakers of that language, then the demands of the exam should be those involved in 'communicating with a native speaker'. The point, once again, is the difference between experts and novices, but in a rather different way this time. We have some evidence (Pollitt & Murray, 1993) that judges with no experience of teaching look for different criteria than trained judges look for, and that they may be quite happy to ignore errors and hesitations (for example) if these do not impede understanding. Teachers may be biased by their professional experience to pay too much attention to the elements that they are used to thinking about explicitly in the teaching context. There may be many other cases, especially in vocational assessment, where this sort of 'consumer comparability' would be worthwhile. If the hairdressers' customers go away equally happy then, by definition, the standards of the hairdressers are equally high.

Reference

Pollitt, A., & Murray, N.L. (1993). What raters really pay attention to. Language Testing Research Colloquium, Cambridge. Reprinted in M. Milanovic & N. Saville (Eds.), (1996), *Studies in language testing 3: Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.