Cambridge Exam Research
www.camexam.co.uk

Evaluating the evidence in assessment

Ayesha Ahmed and Alastair Pollitt

A paper for the 34th Annual Conference of the

International Association for Educational Assessment

Cambridge, September 2008

Evaluating the evidence in assessment

Ayesha Ahmed and Alastair Pollitt

Cambridge Exam Research

Abstract

What is our warrant for saying "Student X deserves a Grade C" ? It must be based on evidence, and the only evidence we see is what students produce during the exam. For valid assessment two criteria must be met: the examination must elicit proper evidence of the trait, and we must evaluate the evidence properly.

This highlights the importance of ensuring quality in the *mark schemes* with which we evaluate the evidence as well as in the *questions* which elicit it.

Our recent research shows that improving mark schemes can make more impact on validity than further work on improving questions.

In this paper we will outline a procedural model for maximising construct validity: at its heart is the concept of Outcome Space, the range of evidence that students produce.

The model aims to ensure that our mark schemes evaluate this evidence properly in terms of the achievement trait we want to assess. This model has been developed in consultation with senior examiners and exam board personnel. A key part of it is a taxonomy for classifying mark schemes and evaluating their suitability for different types of questions. We will present examples to illustrate the essential elements of an ideal mark scheme.

Validity and Marking

We begin with the concept of validity. Traditionally defined as the extent to which a test measures what it is intended to measure, the concept has recently been redefined to emphasise that it is the use of tests, or the inferences that are made as a result of using them, that should be deemed valid or otherwise (Messick, 1989). Nevertheless, for the purpose of examiners constructing examinations, the older idea of *construct validity* is still what matters most. Indeed, in the current *Standards for educational and psychological testing* the word 'construct' is considered redundant, and the notion of 'construct validity' is equated to validity, since "all test scores are viewed as measures of some construct' (AERA et al, 1999: p174).

However validity is conceived, it is obvious that for an exam to lead to valid assessment the students who are 'better' at the subject must somehow get more marks than the others. In the first half of this paper we consider what it means to be 'better', and how examiners can get evidence that shows who is 'better'. In the second half we will focus on how to ensure that the 'better' students get more marks.

The trait we want to assess

The starting point must be a clear consensus amongst the examiners of what it means to be 'good' at the subject. Traditionally, this comes from the Aims and Objectives of the syllabus, which spell out what students are meant to achieve during their course of study. There is, however, a problem with this. The Aims and Objectives are primarily written for teachers rather than for examiners, to describe to them the things they are meant to instil in their pupils. Aims tend to be presented as a list rather than as a coherent description of 'goodness' that could guide examiners in writing questions and mark schemes. The following example from a current syllabus for GCSE Geography is typical.

Aims

The aims set out below describe the educational purposes of following a course based on this specification. Some of these aims are reflected in the assessment objectives, others are not readily translated into measurable objectives. They are not listed in order of priority.

This specification offers opportunities for students to:

a. acquire knowledge and understanding of a range of places, environments and geographical patterns at a range of scales from local to global, as well as an understanding of the physical and human processes, including decision-making, which affect their development;

b. develop a sense of place and an appreciation of the environment, as well as awareness of the ways in which people and environments interact, the importance of sustainable development in those interactions, and the opportunities, challenges and constraints that face people in different places;

c. develop an understanding of global citizenship and the ways in which places and environments are interdependent;

d. appreciate that the study of geography is dynamic, not only because places, geographical features, patterns and issues change, but also because new ideas and methods lead to new interpretations;

e. understand the significance and efforts of people's values and attitudes, including their own, in how decisions are made about the use and management of environments and resources, in relation to geographical issues and questions;

f. acquire and apply the skills and techniques – including those of mapwork, fieldwork and information and communication technology (ICT) – needed to conduct geographical study and enquiry.

Geography A - AQA GCSE Specification, 2010

A more integrated approach is taken by QCA in describing the "importance" of each subject in England's national curriculum:

The importance of geography

The study of geography stimulates an interest in and a sense of wonder about places. It helps young people make sense of a complex and dynamically changing world. It explains where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies and environments are interconnected. It builds on pupils' own experiences to investigate places at all scales, from the personal to the global.

Geographical enquiry encourages questioning, investigation and critical thinking about issues affecting the world and people's lives, now and in the future. Fieldwork is an essential element of this. Pupils learn to think spatially and use maps, visual images and new technologies, including geographical information systems (GIS), to obtain, present and analyse information. Geography inspires pupils to become global citizens by exploring their own place in the world, their values and their responsibilities to other people, to the environment and to the sustainability of the planet.

QCA: National Curriculum – Geography key stage 3

It should not be difficult to rewrite these *Aims* into a form that examiners will find useful, expressing clearly and simply the essential qualities that they should look for to identify 'good' students.

Evidence and Validity

Awarding a mark or grade to a student must be done on the basis of evidence, and the only evidence that examiners have is the performance the student produced in the exam. It is therefore essential that the exam elicit the right sort of evidence, and that the mark scheme evaluate this evidence fairly. Although exams in England sometimes demand oral, artistic or other performances, we are most often dealing with a written performance as evidence.

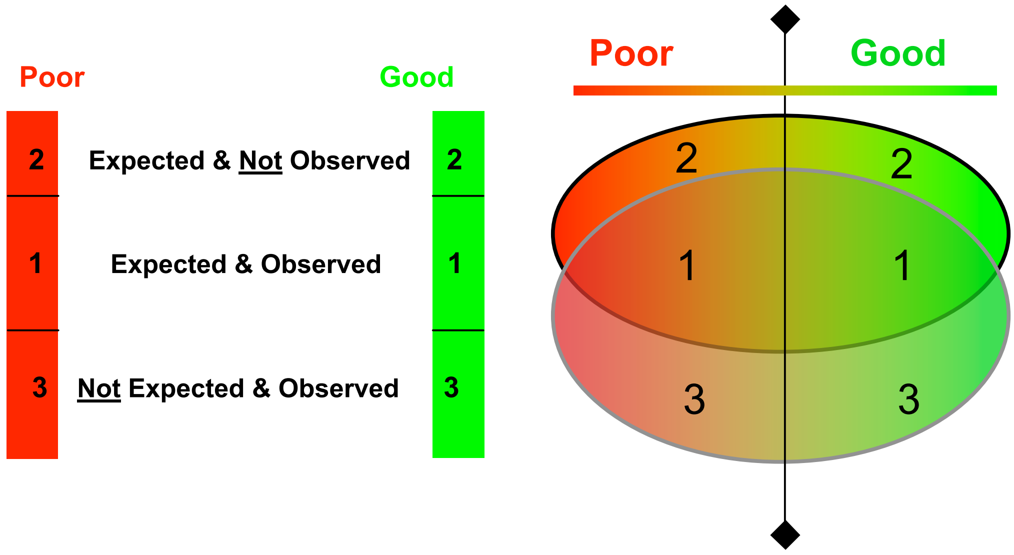
So how do we ensure that the evidence being produced by students is the right sort for us to use to grade them? The answer to this lies in designing exam tasks that represent the trait we want to assess, as described in the Aims or Importance statements: specifically, tasks that ensure the students' minds are doing the things we want them to show us they can do. This will generate evidence we can then use to grade them fairly on their knowledge, skills and understanding of the trait. (Pollitt & Ahmed, 1999; Pollitt et al, 2008)

However, a valid exam that stimulates students into producing the right evidence is not enough. We also need a tool for evaluating the evidence, that is a mark scheme. To explore the issue of mark schemes further it is useful to consider the concept of Outcome Space.

Outcome Space

The concept of outcome space was introduced by Marton & Saljö (1976), in a study of how students read an academic article, to describe the range of responses the students made when asked questions about the article. Their interest was in qualitative differences in how students responded that might indicate differences in their approaches to reading and studying.

The concept can usefully be generalised to any exam question, to represent all of the responses students make to it. For single mark questions these responses may vary in terms of how 'right' or 'wrong' they are, along a continuum from perfectly right to completely wrong; when the response format is more complex, and especially when more than one mark is available, they may also vary in terms of how well the task is completed. To capture both dimensions of correctness and quality of response we illustrate the outcome space as ranging continuously from 'Poor' to 'Good' as shown below.



We will discuss later the significance of the line that divides 'Poor' from 'Good'. For the moment, it is just a convenience for describing the six zones we identify in the outcome space.

Poor 1 and *Good 1* represent the answers the question is meant to elicit and does. Note that it is very important to consider all of the poor answers as well as the good ones, if the question is to succeed in its aim of validly discriminating between poor and good students. For valid assessment we would like these zones to be as large as possible, as they indicate students behaving as the examiners intended.

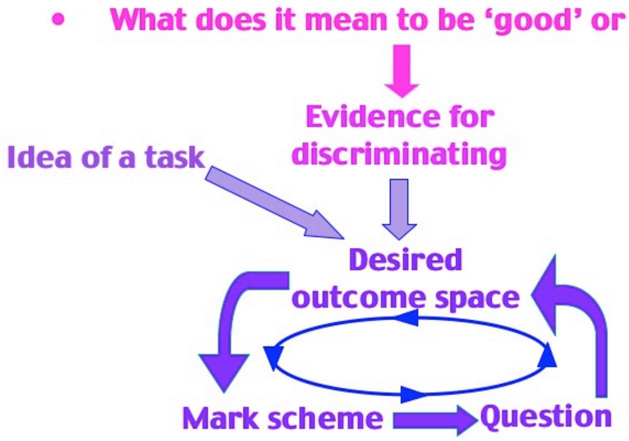
Poor 2 and *Good 2* represent responses that the examiners expected to see but that did not in fact occur; this would include any alternative good but obscure answers, as well as anticipated errors that didn't happen. Some of this space – especially *Poor 2* – is unavoidable, but examiners should at least pause to think why no students came up with errors that the examiners expected them to make. Perhaps the question wording allowed students to avoid these errors?

Poor 3 and *Good 3* are more problematic in terms of validity, as they represent outcomes that were not anticipated by the examiners and cannot, by definition, be included in an initial mark scheme. Any frequently occurring answers in the *Poor 3* zone may indicate a way in which the question could plausibly be misunderstood, an ambiguity or an unfair distraction, and shows that the examiners had lost control of the students' thinking processes. Any response that has to be classified as *Good 3*, an unanticipated but correct response, is more obviously an indication that the examiners had lost control of the question.

The concept of Outcome Space is central to exam construction if we want to provoke valid evidence and assess that evidence validly. We have developed a systematic approach to writing questions and mark schemes in terms of Outcome Space, and we call this approach Outcome Space Control and Assessment (OSCA).

OSCA

When writing a question and mark scheme examiners need to keep the concept of Outcome Space at the forefront of their minds, because if they can control the outcome space then they are in control of the validity of their assessment. Our system can be illustrated diagrammatically as follows:


What does it mean to be 'good' or 'poor'?
Evidence for discriminating
Idea of a task
Desired outcome space
Mark scheme
Question

An essential aspect of the system is that examiners start with the idea of the question and the range of evidence – good and bad – that they want to see, and move from this first to create the mark scheme. Only after a clear view has been formed of the kind of evidence desired and of how it will be evaluated, is the wording of the question finalised. Examiners need to think constantly about the Outcome Space they want to see, and how they want to assess this Outcome Space, i.e. how they want to mark it. Then they will be able to write a question that will generate the evidence of achievement that they want to see.

When we start with the ideas of evidence and Outcome Space, the marking process becomes central. This led us to consider the essential differences between mark schemes in the range that we have seen, and to ask if different types are appropriate for different types of question.

Question types

There are many kinds of task commonly used for assessment, but all of them ask students to provide evidence of their mastery of what is deemed important in the subject. In order to judge this evidence, we find that a crucial factor is the degree to which the students' responses are **constrained** by the task. Consider this ordered list of task types:

Multiple Choice
Selected Response
Cloze/completion
Short Answer Questions
Structured Questions
Structured Essays
Unstructured Essays
Projects
Explorations

In the first few of these, the student has little freedom to choose *how* to respond, and their response will be judged only by how *correct* it is. In the last few the idea of correctness becomes relatively unimportant, and students are assessed mainly for the *quality* of their answers. As the constraint on the response format reduces, the emphasis shifts from *what* students' minds can do to *how well* they can do it.

Mark schemes evaluate evidence – answers which are the output from students' minds. It follows that they must be designed to suit, in each particular case, the nature of the processes going on in the minds that produced these responses.

The Function of a Mark Scheme

We have looked at thousands of mark schemes in the course of our research and our aim here is to define how good ones differ from poor ones; we do this in terms of how well they fit their purpose. What is a mark scheme for?

Look again at the Outcome Space diagram. Central to it is the boundary between Good and Poor responses. A competent marker will have no difficulty recognising excellent responses, and little difficulty with responses that are very wrong, and there is not much need for a mark scheme to identify these. Where the marker needs help is around the boundary: how good does an answer need to be to get the single mark available, or how many of the available marks should a partially right response get? The purpose of a mark scheme is to help markers award marks fairly and consistently, so that the final overall mark a student gets will accurately reflect the quality of the evidence of achievement they have presented. The place where markers need help to achieve this is in the central areas of the outcome space. Our taxonomy is essentially defined, within each category of task types, by how much help markers are given with these crucial decisions.

In our recent report to QCA concerning the quality of GCSE papers (Pollitt, Ahmed et al, 2008) we concluded:

"As a priority, training in how to write mark schemes will probably lead to more immediate improvement in exam validity than any other measure."

It is to this end that we are now presenting the taxonomy of mark schemes. It is based on a survey of Scottish questions and mark schemes (Pollitt, Walker and McAlpine, 2005) and was developed as part of a study of GCSE papers funded by QCA (Pollitt et al, 2008).

The Taxonomy

The nature of the help markers need differs for different types of task. In the simplest category – Very Constrained questions – the function of the mark scheme is to define the boundary between *Correct and Wrong*, or between scores of 1 and 0. For Semi-Constrained questions the range of responses is greater and the function of the mark scheme shifts to helping markers judge whether a particular response shows *enough evidence of Correctness or Goodness* to merit a mark, and perhaps to help them decide how many of the available marks it should be given. With Un-Constrained questions, the concept of correctness may fade almost completely away, and the function of the mark scheme becomes mainly to help markers rate the *Quality* of the responses they see.

In presenting this taxonomy we might give the impression that every mark scheme should aspire to the top level of our system, but this is not necessary. In many cases, a 'lesser' mark scheme may be quite adequate. We do think, however, that it is better to err on the side of too much rather than too little help for the markers.

How the taxonomy classifies VC mark schemes

We begin with the simplest question type, where students have very little freedom to respond in unexpected ways. How does the principle of helping markers to award partially correct responses apply here? Our scheme is:

VC.3	**Rule/principle to differentiate answers**
VC.2	**List of right + list of wrong** • **Examples** • **Complete**
VC.1	**Complete list of right answers**
VC.0	**No guidance / Model answer**

VC.0

Type VC.0 is never recommended. Schemes of this kind give markers no help at all. Consider the following example:

(g) Complete the following sentence about private sector businesses:

The capital of a private business is contributed by………….……

…………………………………………………………………...…. **[1]**

MS:

The capital of a private business is contributed by **the owners/shareholders**

In this case the examiner has decided that the answer must be 'owners/shareholders' . This gives no guidance to markers on how to deal with answers such as 'investors', 'capitalists', 'financiers', 'banks', 'buying shares', 'investment', 'private individuals', 'people who hope to make a profit'. Should any of these answers get the mark? Are they close enough to the model answer or not?

VC.1

Type VC.1 is a complete list of right answers, and for Very Constrained questions this is sometimes adequate. However, it is always better to consider certain of the possible wrong answers as well i.e. type VC.2. Looking at the example below, it seems that a complete list of right answers is good enough for a defined set such as directions in geography.

(iii) The location from which Photograph B was taken is shown on Figure 1a. In which direction was the camera pointing?

MS:

Point mark
south; south west; south south west; S; SW; SSW

However, if we consider an answer such as 'towards the river' it is clear that even in a case such as this there may need to be more guidance for markers. The mark scheme would benefit from examples of this sort of likely but 'obviously' wrong answer.

VC.2

Moving up the hierarchy to type VC.2 then, a complete list or a list of examples is given, both for right answers and for wrong answers. The example below is from a mathematics question.

A bag contains 7 blue, 5 green and 3 yellow balls. Work out the probability that when one ball is chosen at random it is

(i) black [1]

The mark scheme consisted of a list of right answers and a list of wrong answers as follows.

> Accept: 0, 0/15, zero, nought, 0%
> Reject: none, impossible, nil

This list is helpful but how should examiners distinguish any unanticipated answers to decide whether they fall into the 'accept' or 'reject' category?

VC.3

Here a principle is defined to separate right from wrong responses. Implicit in this list above is the principle that in order to get the mark students must respond with a zero in the form of a number, and not just a word that also means zero. Making this principle explicit would have provided all that the markers needed in order to award marks to the students' responses fairly. This would then be classified as VC.3.

How the taxonomy classifies SC mark schemes

Moving to Semi-Constrained questions involves loosening the marking process in two ways. First, we need to start to think of responses *qualitatively* as good enough or not good enough for credit, rather than as right or wrong, and the help markers need is to decide what constitutes 'enough'. Further, as a consequence of this shift, examiners often choose to award *partial credit* for answers that are creditable but not perfect. These two changes make it particularly difficult to write good mark schemes for these questions, but these are the most common type of questions in English school-leaving exams.

When marking these we are considering points of content as in the Very Constrained questions, but also points of quality. Our scheme is:

SC.3 Rule/principle to differentiate responses

SC.2 List of good + list of poor
- **Examples**
- **Complete**

SC.1 List of good responses
- **Examples**
- **Complete**

SC.0 No guidance / Model answer

The classification of mark scheme types for Semi-Constrained questions is clearly very similar to the one for Very Constrained questions. However, we refer to good and poor rather than right and wrong answers, as we are thinking about quality as well as correctness

SC.0

The lowest end of the typology, SC.0, may be a model answer as seen below.

(i) What is a quality circle? (1)

MS: (i) a group of employees that meets to identify quality problems, thinks of solutions and makes recommendations

It is not made clear to examiners how much of this model answer is needed to gain the one available mark. Would the response 'a group who identifies problems in quality' get the mark?

SC.1

Again, a better mark scheme will try to list all acceptable answers. The following example shows the shortcomings of such a mark scheme.

1 a iv) This valley has been created by a glacier, which has changed the shape of the land by a process known as glacial **abrasion**.

Explain in detail how this process works. [4]

MS: Ice contains rocks (1) source of rock (1) glacier moves (1) gravity (1) fragments scrape land (1) striations cut (1) surface smoothed (1) analogy (1) process continues through time (1) (Max 4)

There are nine points deemed worth a mark. But consider this response:

> "Gravity makes the glacier move slowly, and over time it smooths the surface."

How many marks does this response deserve? How well does it explain the process? According to the mark scheme it gets 4 out of 4, even though there is no mention of the glacier scraping the land. The problem illustrated here is that a points mark scheme is unable to differentiate the importance of different points – every point is treated as equal in value. There is no reward for selecting the most important points; in fact, it pays to mention everything you can think of, even if it may not be relevant. 'Points' mark schemes like this are very common in exams in England.

SC.2

A better mark scheme for a Semi-Constrained question will list acceptable and unacceptable responses as seen below.

(ii) Describe how the handle could be made more comfortable to hold.

....

.... **(1)**

MS:

(ii) Round off the ends/edges [1]

Padding, fabric, foam, rubber etc – NO mark

The examiners have clearly thought about the outcome space here and considered what gains credit and what does not.

SC.3

To take this up to the final level, the best mark schemes will consider what makes responses good or poor, that is what distinguishes a good response from a poor one. This will result in a statement of a rule or principle for marking the responses, similar to VC.3.

For example, this mark scheme contains a clear basic rule for markers to use to decide on whether to give the mark:

(i) What is meant by a renewable source of energy? (1)

MS: Credit a simple statement.
Bottom line of 'doesn't run out'.
No credit for exemplification. [1]

As does this mark scheme:

(ii) Calais has a warmer winter and a cooler summer than Wroclaw. Explain why.

MS: Looking for answers related to distance from the sea therefore latitude is not credited.
Land heats up quicker than sea (1)
A clear distinction between land and sea heating. (2) (3)

How the taxonomy classifies UC mark schemes

For Un-Constrained questions the marker's task is to judge the quality of the response against some systematic criteria, and the emphasis is often on the quality of the writing as much as, or even more than, on the accuracy of the content (O'Donovan, 2005). In this respect it is interesting to consider the differences between the assessment of writing in a test of language and in an exam of another subject where there is important content other than the language itself.

Until the 1980s language testers generally assessed writing using 'grade descriptors' that described aspects of language form such as grammar, vocabulary, style, or structure. The emergence of 'communicativeness' as a criterion (Canale & Swain, 1980) enabled a new style of assessment of writing (and speaking) in which the content and the effectiveness of the essay were for the first time given serious attention. Today, applied linguists still stress the effectiveness of a piece of writing as the principal criterion for assessment. Note that this requires a marker to blend concern for both language forms and content in whatever balance is most appropriate for the particular task in question: the criteria are always to some degree specific to that task.

Assessment in content-based subjects has followed these linguistic developments to some extent. What are called 'levels' mark schemes have largely replaced 'points' schemes, at least for questions worth more than a few marks, with levels defined for aspects of both language and content, but the final step to assessment for overall 'effectiveness' is often missing. Once again, we see these forms of assessment as giving different degrees of help to markers who must evaluate real pieces of students' writing.

Again our taxonomy recognises four levels:

UC.3	**Specific trait interpretation**
UC.2	**Analytic levels** • **explicit weighting** • **inplicit weighting**
UC.1	**Holistic implicit levels**
UC.0	**Model answer**

UC.0

The Model Answer is not recommended in any circumstances. In our experience, model answers for this sort of task are often the 'ideal' answer that the expert would write, and are better than it is reasonable to expect from real students. No guidance is given as to what is essential in a real answer, and what is merely desirable: no help is given for judging how many marks to give to any particular real answer.

UC.1

Types UC.1 and UC.2 are both levels of response mark schemes. UC.1 is called 'Holistic Implicit' because judges are asked to make an overall assessment of the student's complete performance, but there is no explicit weighting given for the components: it is sometimes referred to as 'best fit'. An example is:

Levels of response mark scheme. Work upwards from lowest level.

Level 1 Choice of case study applied reasonably well. Gives simple description or explanation. Information is communicated by brief statements. 1/2 marks

Level 2 Choice of case study applied well. Gives descriptive points in more detail but little explanation. Communication begins to show structure with occasional use of specialist terms. Sentences show some coherence but occasional errors in spelling, punctuation and grammar. 3/4 marks

Level 3 Appropriate choice of case study applied well. Provides a balanced account which gives descriptive detailed points with some explanation. Communication has structure with some use of specialist terms. Coherent sentences with few errors in spelling, punctuation and grammar 5/6 marks

Level 4 Appropriate choice of case study applied very well. Provides a balanced account which includes specific description and explanation. Communication is logical and includes specialist terms. Spelling, punctuation and grammar have considerable accuracy. 7/8 marks

Total: [30]

Looking at Level 2, markers have six discrete dimensions to consider:

> Case study; Description/explanation; Structure; Use of specialist terms; Sentence coherence; Spelling, punctuation and grammar.

This is the kind of scheme familiar to language testers as ‘holistic scoring’. But how should judges rate responses that use specialist terms well but with poor spelling, punctuation and grammar? What about a response with good application but poor sentence coherence. Or vice versa? It is this implicitness that is the main source of marker unreliability. UC.1 is appropriate when a child must be classified as wholly belonging in just one ‘best fit’ category – as in England’s National Curriculum system, or in a placement test – but it is seldom appropriate for a single question in an exam.

UC.2

Type UC.2 separates the components that are to count, so that markers are judging a response on, for example, Explanation *and* Communication, with each Level described for each component. An implicit version of this still gives no indication of their relative weights, while an explicit version tells markers clearly how many marks to allocate for each dimension. This resolves the problem of how to mark those responses that excel on one area but fall down in another. An example of an explicit UC.2 is:

	AO3 (max 4 marks)	**AO4 (max 4 marks)**
Level 2	Good analysis in context (3-4 marks)	Good judgements offered based on balanced analysis (3-4 marks)
Level 1	Low level analysis/no context (1-2 marks)	Some judgement offered based on analysis (1-2 marks)

Whether explicit or not, this type of scheme is known to language testers as ‘analytic scoring’. As is the custom there, UC.2 schemes tend to be sets of generic descriptors, rather than ones designed to take account of specific features of a question, and markers often comment that they had difficulty deciding how good the student’s ‘analysis’ needs to be to qualify as Level 2.

UC.3

In principle, type UC.3 is the most adequate way to mark a response to an Un-Constrained question. A Specific Trait Interpretation describes levels of response to the task that was actually set, rather than to a generic one, and with reference to the essential trait the examiners intended to assess. For example:

Level 3 5	To reach Level 3 there must be explanation of causes and effects, well linked to a case study
Level 2 4-3	Specific detail of an example must be included to reach level 2. For top of level there should be explanation of either causes or effects and both should be mentioned.
Level 1 2-1	Descriptive comments about causes and/or effects of cliff recession.

To score highly on this task students must show evidence of what is deemed important in Geography – they will be ranked in order of their ability to “*explain … how places and landforms are formed*”, to quote from the QCA ‘Importance’ statement given earlier.

There are several examples of UC.3 scoring systems in language testing, including *primary trait scoring* as used in the USA's National Assessment of Educational Progress (Mullis, 1983).

Conclusions

The purpose of our taxonomy is to show how mark schemes may be designed, or improved, to enhance the validity of assessment. It is based on two premises: that a mark scheme is intended to help markers decide how many marks to award each of the actual responses that they see from real students, and that they should award these marks in accordance with a consensual view of the trait they want students to demonstrate. To meet the first of these, a mark scheme must concentrate on helping markers score responses in the middle ground; the very good and the very poor are easy, but giving proper reward to those that are intermediate is much harder. To meet the second, this guidance must be based explicitly on how the particular task will demonstrate students' degree of mastery of what is important in learning the subject.

If the marks each student gets are to be valid indicators of achievement, then the task must provoke the desired kinds of mental behaviour in their minds. The students' minds must be behaving as intended by the examiners, so that we can elicit evidence of achievement.

The other essential step is to evaluate the outcomes with mark schemes which reward relevant evidence – and only relevant evidence – of the trait we seek. This all-important evidence is what we describe as the *desired* Outcome Space.

In this conception of constructing examinations it is the elicitation and evaluation of the Outcome Space that is the key to valid assessment. The writing of exam questions and the writing of mark schemes are equally important as the two features of the system we call Outcome Space Control and Assessment.

References

AERA, APA & NCME (1999). American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *The standards for educational and psychological testing*. Washington, DC.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1-47.

Marton, F & Saljo, R, (1976) On qualitative differences in learning: 1 - Outcome and Process. *British Journal of Educational Psychology*, 46, 4-11.

Messick, S. (1989) "Validity," in *Educational Measurement*, ed. R. L. Linn. New York: Macmillan, pp 3-103.

Mullis, I.V.S. and Denver, CO (1980). *Using the primary trait system for evaluating writing*. National Assessment of Educational Progress, Education Commission of the States.

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.

Pollitt, A & Ahmed, A (1999) *A new model of the question answering process*. IAEA, Bled.

Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J and Davidson, M (2008) Improving the quality of GCSE Assessment. London: QCA. http://www.qca.org.uk/qca_15954.aspx

Pollitt, A., Walker, H. and McAlpine, M. (2005) *Final report on the description of question and marking models - External assessment model*. Report to SQA.