*Cambridge Exam Research*
*www.camexam.co.uk*

# Outcome Space Control and Assessment

**Alastair Pollitt and Ayesha Ahmed**

**A paper for the 9th Annual Conference of the**

**Association for Educational Assessment – Europe**

**Hissar, November 2008**

## Outcome Space Control and Assessment

## The cognitive validity of questions and mark schemes

For several years we have studied exam questions, with the aim of improving the validity of examinations (Pollitt & Ahmed, 1999, 2000, 2001; Ahmed & Pollitt, 2001, 2003). Although it is common to talk as if students' achievement is measured by the examination, it is the question that does the measuring. A student's grade is determined by their test score, which is nothing more than the sum of their scores on the questions. If they are not given an appropriate score on any question - a score that accurately reflects their achievement - then invalidity is introduced that cannot later be overcome. Our focus, therefore, has been on the validity of exam questions.

We argue that:

*A question can only contribute to valid assessment:*
*if the students' minds are doing the things we want them to show us they can do.*

More recently, after scrutinizing several thousand questions and mark schemes (Pollitt et al, 2008), we have come to look more closely the mark schemes since these determine the student's score on the questions. Now when we talk about the exam question or task we are actually referring to the whole 'question + mark scheme'. We therefore expand our statement about validity:

*An exam task can only contribute to valid assessment:*
*if the students' minds are doing the things we want them to show us they can do;*
*and if we give credit for, and only for, evidence that shows us they can do it.*

To unpack this statement we must consider what determines how many marks a student gets on a question. This depends on the kind of question. Some consist of one or more quite discrete tasks, each with its own specific sources of difficulty, where a student will get each mark if they can satisfy the demands of each sub-task. In less constrained questions students will be given more or fewer marks according to how well they perform, but again *how well* is defined in terms of the demands set in the question by the examiners.

It is clear from this that the validity of a question depends on two factors. First the question must set appropriate demands to the students, so that their minds are required to engage in appropriate mental activities as they answer the question. Second, their success in these appropriate activities must be what determines their score on the question. Our definition of validity is therefore couched in terms of cognitive activity in the students' minds:

*An exam task can only contribute to valid assessment:*
*if the students' minds are doing the things we want them to show us they can do;*
*and if we give credit for, and only for, evidence that shows us they can do it.*

This leads us to the concept of evidence. In order to give students a score the question must elicit appropriate evidence from them and the mark scheme must evaluate fairly the evidence they produce.

## Evidence: evaluation and elicitation

Let's summarise the argument so far. For each question in an exam:

1 the job of the examiners is to elicit and evaluate valid evidence about students' achievement;

 © CamExam, 2008

2 the job of the mark scheme is to evaluate that evidence fairly, in accordance with the trait that the exam is intended to measure;

3 the job of the task set by the question is to elicit that valid evidence from the students.

## The cognitive validity of an examination

### Valid questions do not necessarily make a valid exam

If it is the question – note that we always include the mark scheme as part of the question – that does the measuring, then what is the role of the test? We can add to the summary above:

4 the job of the test is to ensure that the total set of questions samples appropriately the whole domain that is meant to be assessed.

It is quite possible for every question to be valid, in that it requires students to do some of *the things we want them to show us they can do*, but for the set as a whole to fail to cover the domain adequately. A key issue for examiners, then, is to determine how best to ensure adequate coverage. How should this be done? What does it mean 'to sample the domain'?

We begin by extending our cognitive definition of validity from the question to the whole test:

*An exam can only contribute to valid assessment:*
*if the students' minds are doing a representative sample of the things we want them to show us they can do;*
*and if we give credit for, and only for, evidence that shows us they can do it.*

## Mental behaviours

This definition reminds us that an examination is fundamentally a form of mental measurement. Validity can only come from ensuring that, over the whole set of questions each student attempts, they provide adequate evidence of how well they can do the *mental* things that the examiners want them to show that they can do.

Since it is mental things that we want to see them doing, what matters is sampling mental behaviours. How do we do this? How can we describe a subject in terms of mental behaviours?

The traditional ways to detail a subject are in terms of *skills* and *content*.

### Skills and Content

There is a problem with each of these. Skills are a kind of abstraction from behaviours. As a general description of the kinds of thinking that students will be expected to engage in they may serve a useful purpose in the exam syllabus, but they are less helpful when used in the specification for an exam. The abstraction from behaviours to skills is problematic: typically, in British exams at least, a simplified version of Bloom's taxonomy (Bloom et al, 1956) is used, but the assigning of particular questions to the categories is often debatable. Test specifications are used, not as aids to ensure test validity, but as rules to enforce superficial equivalence between different test forms. Arguments are common at meetings of the Question Paper Evaluation Committees, and often concern manipulating questions to appear to comply with the rules.

 © CamExam, 2008

The use of content in the specification also distracts examiners from attending to validity. The description of content used in British exams is generally a list of topics from the syllabus, and it would be more accurate to describe the approach as an attempt to achieve 'topic validity' than to think of it as addressing content validity. The concern seems to be about avoiding accusations of unfairness by making different test forms look similar, rather than anything to do with validity.

If a proper definition of 'content validity' were used, examiners' attention would turn, as we suggest, to the behaviours required of students. If 'content coverage' is one aspect of content validity, the other is 'content relevance' which 'requires the specification of the behavioral domain in question and the attendant specification of the task or test domain', Messick (1980, p1017). Even more explicit is this definition used in applied psychology in the USA:

> *The Uniform Guidelines on Employee Selection Procedures (1978) state that "To demonstrate the content validity of a selection procedure, a user should show that the behavior(s) demonstrated in the selection procedure are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job."* *Burns (1995)*

In their summary of many psychometric definitions of content validity, Haynes, Richard & Kubani (1995) show that, when considered from a psychological perspective, true content validity is indistinguishable from construct validity:

> *Although worded differently, most of these definitions encompass concepts embodied in the following definition:*
> *Content validity is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose.* *(p2)*

Can we ground our sampling more directly in behaviour?

## Performance assessment

Of course it seems different in subjects such as drama and music that involve what is called Performance Assessment ('PA'). Here it is obvious that behaviour is being assessed. But it is also true that every response a student makes in any exam is a 'performance' – most often a written performance. All we can ever 'see' is this performance, from which we try to infer the student's competence.

The difference inherent in PA seems to be that the performance that we observe is considered to be a reasonable facsimile of the performances we could (in principle) observe in 'real life' beyond the classroom or exam hall. In language testing, this approach has been developed most fully by Bachman (eg Bachman & Palmer, 1996) who aims to describe a domain of situations and functions from which an exam in, say, speaking can be constructed by random sampling. Weir (2004) provides a more detailed framework for analysing the components that determine the nature of a test task; he defines *contextual validity* as: "...the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (p19).

This is broadly the attitude to validity that we are advocating, but the descriptions by Weir and Bachman are more sociolinguistic than psycholinguistic, with many referring to such factors as audience, setting or topic. We would separate these factors that describe the context within which students are asked to perform from those that determine what kinds of things their minds are being asked to do in the performance.

The latter are what we call demands. They are essential to the performance, in that an exam that gets them wrong cannot be valid, since the students' minds will not be doing the things we want them to show us they can do.

## Demands and cognitive validity

Demands are an essential link in making this inference from performance to competence. In order to see how we can assess the students' mental behaviour we need to consider what behaviours the task demands from them. Demands are what the examiner puts into a task. The ones that the examiner intends to be there are construct relevant (CR) demands; those that slip in to the question unintentionally are construct irrelevant (CI). For example, questions with lots of reading in a science exam may be considered to contain CI reading demands, while a reading task that can only be understood by those who are familiar with scientific concepts might contain CI demands on scientific knowledge. Demands are there in the question from the start, before the student even meets the question. They determine what it is that students' minds will be doing in an exam, and will determine how validly we can make that inference of competence from a given performance (see Pollitt, Crisp & Ahmed (2007) for a more detailed discussion of demands).

Difficulty however is not defined until the students have taken the test. Difficulty is simply the empirical measure of how successful students were at scoring marks on the task. It depends on the mark scheme as much as on the question, since a more 'generous' mark scheme really will make a question easier, even though its demands do not change.

What are the demands we want to see students dealing with? If we understand what demands we are putting into a question, then we will be in control of the mental behaviour that occurs when a student meets a question. If we can control this mental behaviour then we are controlling the students' actual performances. Being in control like this means knowing that the student's performance will provide the right sort of evidence of the trait we are trying to measure. If we get the right sort of evidence then we can measure the students fairly – that is, we can make valid inferences about their competence.

Recently, we proposed that examination syllabuses should specify in advance the range and the levels of demands they intend to make of students (Pollitt, Crisp & Ahmed, 2007). Let us consider this proposal in a little more detail.

Take reading demand first. If a test is meant to be a reading test, then this demand is the crucial CR one, and the syllabus or test manual should describe this in detail. For some other exams, especially such as literature, history, or politics, it may be CR to require students to read particular texts because these are the objects of study, and a student cannot be considered to have mastered the subject if they cannot understand these texts to a certain depth. Apart from these two types of test, reading is CI, and should not interfere with the testing of the intended CR demands of 'doing' science or maths or whatever. The job of the question, after all, is only to communicate the task to the students. Thus, for most school subject exams, the reading that is required of students should be as easy to understand as possible.

A similar analysis can be applied to writing: in some exams the ability to write a well structured argument or explanation is definitely CR – in politics or philosophy, for example, it might be seen as the essence of the trait being assessed. In some others it is important as an indicator of mastery, but in many exams the demand that students should always express their understanding in writing adds a serious CI demand that may invalidate the exam as a test of understanding. Rather than let this happen, the syllabus

should specify what part of the assessment will demand lengthy writing, and let the rest use formats that make less writing demand.

We could deal with other demands, such as memory, speed and stress in a similar way.

The remaining essential aspect of demands is cognitive demand, or the demands that an exam makes on students' mental or reasoning abilities. There are two ways of dealing with these – in terms of the cognitive level of the task processes or in terms of the demand of task characteristics. For the first, we find Bloom's taxonomy unsuitable, as its categories do not easily map onto the actual cognitive processes demanded of students in answering the questions. Instead, for most types of question, we favour the single hierarchy approach first proposed by Peel (1971): he classified the thinking required into the hierarchy: *Mentioning*, *Describing*, *Explaining*, *Extended explaining*. This has been developed further by Sutherland (1982) into a more detailed scheme, and more theoretically by Biggs and Collis (1982). With these schemes there is a direct relationship between the command word (verb) in the question and the level in the cognitive hierarchy (at least if the mark scheme is properly written) making them much easier to use than the more abstract taxonomies.

For the second way of handling cognitive demands, rating scales can be used to judge the levels of demand in a question, a paper, or a whole examination. In the scheme we developed (Hughes, Pollitt & Ahmed, 1998), which was based on work by Edwards & Dall'Alba (1981), overall judgements of Complexity, Resource support, Abstractness, and task or response Strategy are made by examiners or other experts. We further suggest (Pollitt, Crisp & Ahmed, 2007) that it would be helpful if the writers of a syllabus were to specify in advance the profile of these demands that they intend to build into their exams, allowing examiners – and other interested parties – to assess the actual exam against the intentions: in other words, to assess its cognitive validity.

If this approach is adopted, the examiners will necessarily keep at the front of their minds the fundamental issue for validity – ensuring that the tasks are assessing the essence of the trait, the demands that the syllabus says should determine the students' performances.

Given that, it seems less important to specify how the 'topics' in the syllabus should be sampled. If a student can demonstrate the ability to perform at a high enough level in some topics, then it is not very important to establish that they can do so in every topic – in history, for example, it is assumed quite reasonably that a student who is 'good' at Mediaeval Europe would also be 'good' at 19th Century America if they just learned the relevant content. The same is true to some degree of every subject, for the content studied is never all that could have been studied, and the grade a student gets can never guarantee that they have mastered the whole domain.

Demands are more essential than either skills or content, in that they are what defines the trait. The essential point is that, by specifying how the demands are to be allowed to influence performance the examiners are giving themselves a more useful criterion for validity than they can get from any two-dimensional grid of 'skills' and 'content'.

As stated above:

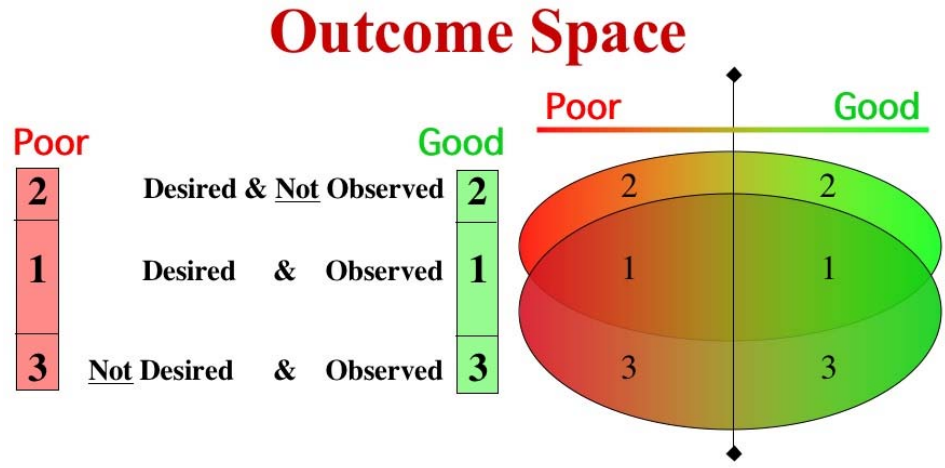
4 the job of the test is to ensure that the total set of questions samples appropriately the whole domain that is meant to be assessed.

This can only be addressed by ensuring that the demands in the total set of questions that constitute an examination form a representative sample of the trait the examiners intended to assess.

## OSCA

We have brought all of the above ideas together into a system for question writers. It describes how we can use the concepts of evidence and demands in order to produce questions and mark schemes that will contribute to valid assessment. Our system is called Outcome Space Control and Assessment (OSCA) and will be outlined below. First it is necessary to consider the idea of Outcome Space. Outcome Space simply refers to the evidence – the responses examiners hope to get as evidence of understanding of the trait, and the responses the students produce.

The concept of Outcome Space (OS) comes from work by Marton and Saljo (1976) in which they used the term to describe the range of responses students produced when asked questions on an academic article. They used OS to describe the qualitative differences in responses. We are using the term in a more general way to refer to exam question responses. For any exam question there will be a range of responses from poor to good responses that students will produce. The diagram below shows how the OS can be divided up:

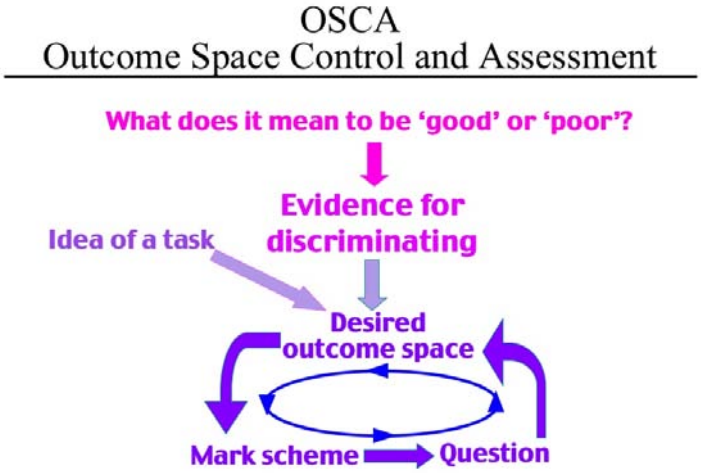



The areas *Poor 1* and *Good 1* represent the answers the question is meant to elicit and does. Note that it is very important to consider all of the poor answers as well as the good ones, if the question is to succeed in its aim of validly discriminating between poor and good students. For valid assessment 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.

We use the idea of OS now to describe how exam tasks should be written in a systematic way in order to maximize validity.




As shown in the diagram, the process of question writing must begin with a shared understanding of the trait examiners are trying to measure, and what it means to be 'good' or 'poor' on that trait. From this understanding, and with an idea of a task in mind, they will then be able to decide on what evidence they would like to see that will help them to discriminate between good and poor performances on the task. This evidence, i.e. these performances, is called the Desired Outcome Space. Next they should consider how they intend to make inferences about the students' competence from their performances, which means they should draft the mark scheme. The crucial issue is how to infer which students have greater competence on this trait – the mark scheme, if it is based on a careful consideration of the evidence, the desired Outcome Space, should facilitate this. Only then do the examiners consider how to elicit the performances they want by working on the precise wording of the question.

There then follows a process of iteration around question, desired outcome space and mark scheme, until the examiners arrive at a question and mark scheme that will elicit the desired outcome space and evaluate it appropriately.

We believe that following the systematic procedure outlined in OSCA is the best way to ensure valid assessment. This is achieved by eliciting evidence of the right kinds of mental behaviour - *the things we want them to show us they can do* – and by evaluating the resulting performances in order to make valid inferences about competence.

## Conclusion

Outcome Space Control and Assessment (OSCA) is a systematised procedure intended to maximise the validity of examination questions. The focus is on the evidence on which assessment is based, and on how best to elicit it and evaluate it.

We have argued that the evidence is actually the mental behaviours of students, and that therefore it is these behaviours that we should concentrate on in designing the questions, the mark schemes and the tests as a whole.

The key to designing appropriate assessments is understanding the kinds and the levels of demands that are appropriate for that particular examination.

At question level this means ensuring that the demands that are put into each question are as Construct-Relevant as possible, and that it is these CR demands that are credited by the

mark scheme. To the extent that this is successful, the questions will each measure some aspects of the trait that the exam aims to assess.

At test level the requirement is that the set of demands assessed by the set of questions constitutes an appropriate representation of the trait. This can be ensured only if the examination syllabus specifies the trait in terms of the demands that successful students are expected to be able to meet. We suggest that the syllabus should contain a list of demands and indicate the role that each demand should play in the assessment – that is, how much of it is appropriate and Construct-Relevant for that particular assessment. Draft papers should then be judged against this specification.

The test specifications commonly used in British examinations seem to us too detailed to help test constructors. They focus mostly on ensuring that different forms of the examination cover the range of possible topics to a similar degree, and that Bloomian 'skills' are similarly represented. But it is demands and cognitive processes that define the trait, and these should take precedence over content or abstract notions of skills. We believe the partly qualitative and partly quantitative kind of holistic specification we advocate would do more to guarantee validity than the over-prescriptive and slightly trivial specifications now in use.

The demands examiners put into an exam paper determine what mental behaviours the students will apply their minds to, and so determine the nature of the evidence available to the markers, and thus the nature of the trait actually being measured. Only through explicit attention to demands can we ensure that the exam does constitute a valid sample of student behaviours.

## References

Ahmed, A & Pollitt, A (2003*) Why do students get exam questions wrong?* International Association for Educational Assessment Conference, Manchester.

Ahmed, A & Pollitt, A (2001). *Science or Reading?: how students think when answering TIMSS questions*. IAEA, Rio de Janeiro.

Bachman, LF and Palmer, AS (1996) *Language testing in practice: designing and developing useful language tests.* Oxford : Oxford University Press

Biggs, JB and Collis, KF (1982) *Evaluating the quality of learning: The SOLO taxonomy*. New York and London: Academic Press.

Burns, WC (1995) *Content Validity, Face Validity, and Quantitative Face Validity*. San Francisco: WCB&Associates. Available at http://www.burns.com/wcbcontval.htm

Bloom, BS, Engelhardt, MD, Furst, E J, Hill, WH and Krathwohl, D (1956) *Taxonomy of educational objectives: The cognitive domain.* New York: McKay.

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.

Haynes, SN, Richard, DCS & Kubani, ES (1995) Content Validity in Psychological Assessment:A Functional Approach to Concepts and Methods. *Psychological Assessment*, 7, 238-247

Hughes, S, Pollitt, A, & Ahmed, A (1998) *The development of a tool for gauging the demands of GCSE and A Level exams questions*. British Educational Research Association conference, Belfast.

Marton, F and 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.

Peel, EA (1971) *The nature of adolescent judgment*. London, Staples Press.

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

Pollitt, A & Ahmed, A (2000) *Comprehension Failures in Educational Assessment*. European Conference on Educational Research, Edinburgh.

Pollitt, A & Ahmed, A (2001). *Understanding students' minds: how to write more valid questions*. Association for Educational Assessment-Europe Conference, Krakow.

Pollitt, A & Ahmed, A, Baird, J-A, Tognolini, J and Davidson, M (2008) *A study to investigate ways to improve written GCSE examinations.* London: QCA. Available at http://www.qca.org.uk/qca_15954.aspx, or at http://www.camexam.co.uk

Pollitt, A, Crisp, V & Ahmed, A (2007) Techniques for exploring the demands of syllabuses and question papers. In Newton, P, Baird, J, Patrick, H, Goldstein, H, Timms, P & Wood, A (Eds) *Techniques for monitoring the comparability of examination standards*. Oxford: Oxford University Press.

Weir, CJ (2005). Language Testing and Validation. Basingstoke: Palgrave Macmillan Ltd