Cambridge Exam Research

University of Bristol

Pearson Research & Assessment

Improving the Quality of GCSE Assessment

Alastair Pollitt & Ayesha Ahmed

Cambridge Exam Research

Jo-Anne Baird

University of Bristol

Jim Tognolini & Michelle Davidson

Pearson Research & Assessment

January, 2008

QCA

This report was commissioned by the Qualifications and Curriculum Authority (QCA).
The content of the report represents the views and conclusions of the author rather than those of QCA.

Contacting the authors

The authors of this report will be interested to hear from readers. Research ideas and their applications are never finalised, but always in the process of being developed, and we will be happy to take into account any experiences or opinions that readers are willing to share with us.

To contact us please use one of these addresses:

alastair@camexam.co.uk   ayesha@camexam.co.uk

We intend to publicise further developments on the Camexam website: www.camexam.co.uk

Summary

Introduction

Our aim is to improve the quality of written GCSE assessment, focusing specifically on Geography, Business Studies and Design and Technology. Our strategy has been to study the examination materials to discover:

- what examiners want the students to show them they know and can do;
- whether these types of question seem likely to elicit appropriate evidence;
- whether the mark schemes are likely to reward students in a way that will lead to valid inferences.

Question validity -– what is a question for?

The purpose of an exam question is to elicit the evidence that the students' minds can do the things we want them to show us they can do. Exam questions are our way of probing students' minds to see how well they have learned the knowledge and skills the course was meant to give them. The job of the question is to direct the student's mind towards doing the things we want evidence about. In a good question the evidence we see will accurately indicate how well students have learned the relevant knowledge and skills.

Question validity – what is a mark scheme for?

It follows too that the purpose of a mark scheme is to ensure that students are given credit when, and only when, they show evidence that they can do the things we want them to show us they can do. If marks are awarded for things that are not evidence of learning, or not awarded for what is evidence of learning, then the scores students get cannot be trusted as indicators of their levels of success.

Method

We analysed questions and mark schemes from one syllabus in Geography, Business Studies and Design and Technology from each of five examination boards. The initial analysis involved coding questions in four categories: command word, response type, marking type and number of marks. This allowed us to look at the different command words and how they were used, as well as the variety of marking methods and response types for each subject. We analysed a total of 1,913 items in this way. At this stage we also flagged questions for the more detailed analysis described below.

Outcome space theory

Marton & Saljo (1976) introduced the concept of outcome space, in a study of how students approached the task of reading an academic article. We have extended the concept of outcome space to more constrained question types, in order to understand what is required of a good mark scheme. A question's Outcome Space is the set of all responses to it, both actual and potential. The point is that, using the mark scheme they are given, a marker has to be able to evaluate every part of that space, every likely answer.

The set of all responses to a question can be separated into six subsets:

Good 1	The subset of	'Good' answers	Observed	Expected
Poor 1	The subset of	'Poor' answers	Observed	Expected
Good 2	The subset of	'Good' answers	**not** Observed	Expected
Poor 2	The subset of	'Poor' answers	**not** Observed	Expected
Good 3	The subset of	'Good' answers	Observed	**not** Expected
Poor 3	The subset of 'Poor' answers		Observed	not Expected

where 'Good' includes responses judged either correct or of high quality, and 'Poor' includes both wrong and low quality responses.

We want the overlap between the expected and observed outcome spaces to be as large as possible, and the rest to be small. In particular we don't want to see many right answers that have not been anticipated.

The Outcome Space Generator

This is a practical tool developed from the concept of outcome space. The idea is to use cognitive psychology, psycholinguistics, and knowledge of how students think and behave in exams to predict the range of answers and answer types that a particular set of candidates might offer to a specific question. A careful description of the Outcome Space will demonstrate any weaknesses in the question or failures in the mark scheme.

To structure the process of Outcome Space prediction, we used a simple three-phase model adapted from a full model of the question answering process (Pollitt & Ahmed, 1999):

1	Reading -	the construction of a mental representation of the task
2	Thinking -	the activation of concepts related to the question words, secondary activation of other concepts, making connections, getting an idea of an answer
3	Writing -	producing a visible response to complete the task

(This is a simplified model as in practice the phases are not discrete).

Findings

Our task was to be critical: to look for problems in the examinations that might indicate where improvements were possible. In studying almost 2000 items – questions and sub-questions – we found problems that could be grouped under four headings:

1. Problems with controlling students' thinking
2. Problems with mark schemes
3. Mismatch of mark scheme and question
4. Subject specific issues

Under (1) we deal with problems where the way that the question was asked may have caused it to fail to provoke the kinds of thinking the examiners wanted. Examples of this include the effects of context, misleading command words, misuse of emphasis, or ambiguity.
Under (2) the mark schemes were sometimes inadequate, with little useful guidance given to markers. In some cases model answers were given which gave them no help at all in evaluating real responses.
Under (3) there was a mismatch between the task set by the question and the way in which credit was awarded. Sometimes points mark schemes were used when a generic scheme would have been better, or the command words in the question would encourage students to give answers which were not rewarded in the mark scheme.

We classified the problems we saw into 25 subcategories under the above headings. The report contains about sixty examples to illustrate the problems.

Another outcome of the initial phase was a scheme for classifying the questions into three types: Very Constrained questions (VC), of which the extreme type is multiple choice, UnConstrained questions (UC), typified by the essay, and Semi-Constrained (SC) questions, in which students are given some but not all the necessary structure for their response. In the GCSE exams that we looked at most questions belong to the SC category, with written responses ranging from a phrase to several sentences, or instructions to draw or modify a diagram. SC questions are the most problematic to mark. We have developed a taxonomy of mark schemes in which we identify the types of mark schemes that are appropriate for questions with different levels of constraint.

How to write an exam question

The best approach to dealing with all of these problems is to adopt a systematic procedure that will support the creative activity of writing questions and mark schemes. In the system that we propose, traditional practice is reversed. The sequence we suggest is as follows:

idea of task → desired outcome space → mark scheme → question

to ensure that the question writers begin with, and always keep in mind, that the purpose of the question and mark scheme is to elicit valid evidence of achievement. The importance of using the correct command word has become clear from this work and we suggest that this should be addressed in the final phase of question writing, with direct reference to the mark scheme that has already been written. In particular, we found that the command word ‘explain’ has many interpretations and caused a number of problems that we describe in the report.

The two processes of writing questions and of constructing exam papers are logically separate, and should be kept apart if at all possible. Questions will be better if examiners are not trying to fill a specification, but are just trying to write good questions. Also, examiners will only have to write the questions they are good at writing, and any time pressure is removed.

Construct Relevant assessment

To summarise the requirements for good, valid GCSE assessment:

1 We need a way to write questions that ensures that most students’ minds **are** doing the things we want them to show us they can do.

2 We need a system that helps examiners produce mark schemes that help markers assess the kinds of scripts they will meet.

3 We need to ensure a good match between the kind of mark scheme and the kind of question so that we actually give credit for the **evidence that students can do** the things we want them to show us they can do.

An exam question 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 how well they can do it.

Conclusions and Recommendations

1. The purpose of an exam task is to elicit the evidence that the students’ minds can do the things we want them to show us they can do.
2. The purpose of a mark scheme is to ensure that students are given credit when, and only when, they show evidence that they can do the things we want them to show us they can do
3. The purpose of an exam question is to convey to the candidates the task they are required to complete
4. A question’s Outcome Space is the set of all responses to it, both actual and potential, both good and poor. Using the mark scheme they are given, a marker has to be able to evaluate every part of that space, every likely answer.
5. Outcome Space can be predicted by using the phases of the Model of the Question Answering Process to think through how a naïve anxious borderline student will approach a question
6. OSCA – Outcome Space Control for Assessment – If we can predict the Outcome Space, we can try to control it, so that we can assess it fairly

7. Question writing should follow this sequence:
    i. the key idea of the task, then
    ii. the desired outcome space,
    iii. then the mark scheme and then
    iv. the question.
8. The desired outcome space should then be revisited to check whether the question will produce the desired effect.
9. We have created a taxonomy of mark schemes for three different types of question – unconstrained, semi-constrained and very-constrained – which define the principal characteristics of good mark schemes.
10. As a priority, training in how to write mark schemes will probably lead to more immediate improvement in exam validity than will any other measure.
11. Training for examiners on OSCA theory and how to use it is recommended.
12. Training – or education – for senior examiners seems particularly urgent.
13. New question creation systems should be sought that help participants develop both professional and managerial skills in assessment.
14. Steps that can help separate the writing of questions and mark schemes from the compiling of papers should be sought and encouraged; any disincentives, like over-tight specification rules or customs, should be removed.
15. To support the continuing training of examiners, awarding bodies should look for ways to feed item level statistical information back to the writing teams as quickly as possible: for the same reason, question writers or QPEC teams should be asked to forecast the average score on each sub-question.
16. The use of command words and phrases should be systematised, primarily within subject areas. Particular attention should be given to the uses of 'Explain", and the use of 'Give *n* reasons …' should be reconsidered.
17. Empirical research should be commissioned into aspects of how command words and phrases are actually used, and into the effects of varying them.
18. The concept of 'house style' should be reconsidered, to ensure that it serves the purposes of good examining rather than superficial design concerns.
19. The classification of cognitive levels in British examination answers by Peel and Sutherland should be considered as a replacement for the variety of systems currently in use that have evolved from the work of Bloom.
20. The wide range of ways in which case studies are used, especially in Business Studies, should be reconsidered, especially in terms of the impact of pre-release and the internet on assessment validity.

Contents:

1 Introduction

Our aim is to improve the quality of written GCSE assessment.

In consultation with QCA we chose to study examinations in Geography, Business Studies and Design and Technology. From our experience with different awarding bodies in England, Wales, Northern Ireland, Scotland and Australia, and with examiners in many other countries too, we believe that these subjects are particularly difficult to assess well – Geography because of the very wide range of content and skills it contains and the others because they represent new academic disciplines not yet well-embedded in the educational system – and we therefore expected that we would find in them plenty of clear examples to illustrate problems that examiners *in every subject* face whenever they construct and carry out an examination.

The aims of the project were initially set out in the specification in these terms:

SCOPE

> *The study will need to consider and evaluate the validity (including reliability) and effectiveness of current examination papers used across a range of subjects with a view to identifying a) issues with current examinations and b) features that would improve both the assessment and its impact on teaching. The work needs to take into account recent reports on GCSEs, relevant research and developments in other countries.*
>
> *The study should focus on validity; this is intrinsic to the assessments themselves, and so is within the control of the awarding bodies. It is relatively well researched and there is an existing and available literature.*
>
> *The aim of the study is not to identify specific problems with particular assessments; these are properly identified and addressed by awarding bodies in their assessment development and evaluation work. Rather, it is to identify larger-scale issues using particular subjects as the focus of the work. The study should cover:*
>
> - *the range of question types*
> - *the use of command words*
> - *the effectiveness of coverage of assessment objectives*
> - *potential unexpected sources of difficulty*
> - *the use of stimulus material*
> - *whether the issues identified apply to the subject under investigation or more widely*
>
> *Although the study will begin with a review of existing examination practices, the successful contractors will be required to engage in creative thinking about improved GCSE assessment.*
>
> *It is intended that the outcomes of this work should contribute to the development of revised GCSE specifications in 2008.*

We come mainly from a psychological background and see educational assessment as an attempt to measure, as fairly and validly as possible, aspects of students' minds. In general we take as given the aims of an examination syllabus and the broad approach to assessment that is adopted by the examiners; our function is to facilitate them in achieving fair and valid assessment of those aims.

The key to writing good exam questions is to understand how students think during examinations, and so to be able to anticipate how they will react to each question. Examiners begin with a shared set of assumptions about what they want the students to show them that they know and can do, and of what are reasonable levels of the various aspects of cognitive demand they can include. The function of the questions is to provoke these kinds of thinking and to elicit evidence that will support valid inferences about how much they know and can do. Our strategy has been to study the examination materials to discover:

- what examiners want the students to show them they know and can do;
- whether these types of question seem likely to elicit appropriate evidence;
- whether the mark schemes are likely to reward students in a way that will lead to valid inferences.

2 Validity, Quality and Item validity

In recent years a single coherent theoretical concept of assessment validity has gained general acceptance. It argues that assessment will only be valid if those who use its results make appropriate interpretations of those results. There is a danger that this conceptualisation may be seen to take the responsibility for validity out of the hands of those who construct examinations, administer, mark and grade them.

2.1 Current interpretation of Validity

The current consensus is mainly due to Messick (eg 1989, 1990). He emphasised that each of the traditional approaches to validity – content, predictive and construct validity – actually described a kind or kinds of evidence that could contribute to the valid use of test scores. None of them was sufficient on its own to support valid use of the results of assessment:

> "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment. *(Messick, 1990, p1)*

The key issues of test validity are “the meaning, relevance, and utility of scores” (ibid, abstract). It is proper to refer to the older conceptualisations as *facets* of validity, or as principles for obtaining evidence on these issues; only if there is adequate evidence for all of them can the test interpretations and uses be considered sufficiently valid. He did, however, also make it clear that one of the familiar approaches – construct validity – was more fundamental than the others, more necessary for validity in any kind of test use:

> "the construct validity of score meaning is the integrating force that unifies validity into a unitary concept." (ibid, p25)

2.2 Intrinsic validity

With this conceptualisation no individual can guarantee validity. Valid inferences can only be made by users interpreting wisely the outcomes of exams that were wisely graded, wisely marked and wisely constructed, from questions that were wisely created by the question writers. Our concern is for test constructors and question writers, and the most – and least – that can be expected of them is that they should provide examination papers that are capable of supporting valid inferences. To achieve this requires two properties from the exam. First, the questions must require the students to demonstrate appropriate levels of the knowledge and abilities that are agreed to constitute the trait of ‘achievement’ in that subject. This involves a demand both that *each* question should require appropriate skills and also that *altogether* the set of questions should constitute a balanced definition of ‘achievement’. Second, the mark schemes used to assess the students’ responses should ensure that more marks are awarded to students who show evidence of more achievement, by giving better answers or good answers to more demanding questions. If these two criteria are met then the exam will produce results – scores and grades – in which the order of students’ scores will properly correspond to the order of how well they have achieved successful learning of the subject skills.

This corresponds rather closely to the old idea of construct validity, though with a modern cognitive emphasis on the psychological processes underlying the observable evidence of achievement:

> “Possibly most illuminating of all are direct probes and modeling of the processes underlying test responses …, an approach becoming both more accessible and more powerful with continuing developments in cognitive psychology: (ibid, p 16).

With this evidential basis, a test may be fit for a specified purpose (such as measuring learning as described in a particular syllabus) even if its results are later used to make unsupportable inferences by their teachers, job or college selectors, politicians or others. This is all examiners can be expected to provide.

2.3 Question validity – what is a question for?

Learning is not visible to the examiners. If they are to judge how much or how well students have learned then the students must provide visible (or otherwise perceptible) evidence of their learning.

The purpose of an exam question is to elicit the evidence that the students' minds can do the things we want them to show us they can do. Exam questions are our way of probing students' minds to see how well they have learned the knowledge and skills the course was meant to give them. The job of the question is to direct the student's mind towards doing the things we want evidence about. In a good question the evidence we see will accurately indicate how well students have learned the relevant knowledge and skills.

2.4 Question validity – what is a mark scheme for?

It follows too that the purpose of a mark scheme is to ensure that students are given credit when, and only when, they show evidence that they can do the things we want them to show us they can do. If marks are awarded for things that are not evidence of learning, or not awarded for what is evidence of learning, then the scores students get cannot be trusted as indicators of their levels of success.

3 The Model of the Question Answering Process

From earlier research we have developed a psychological model of the processes occurring when students answer exam questions (eg Pollitt & Ahmed, 1999). It was developed initially from the study of five GCSE examinations by analysing the responses made be several hundred students, and was then tested extensively through the experimental manipulation of questions and various forms of protocol study and interview (eg Ahmed, Pollitt & Rose, 1999; Pollitt & Ahmed, 2001; Ahmed & Pollitt, 2007). This is a general model applicable to all subjects and at all levels but we have worked mostly at GCSE. For this project we have used this model as a framework for considering all the issues involved in the writing of valid assessments.

The model of the question answering process is divided into phases. The first phase is Learning, which we label **Phase 0** as it occurs before the examination. The learning phase is what we are trying to measure, so in order to write valid assessments it is important to understand something of how students learn.

Learning is initially represented as an episodic memory of the situation in which a concept was learned (Conway et.al, 1997). Students remember the activity of learning, rather than just what the teacher intended them to learn, and their recollection will include memories of many of the particular incidents and non-cognitive elements that accompanied it. The final result of learning will be a multi-modal mental representation: each concept will be accompanied by many memories arising from the setting in which it was learned – memories of events, feelings, smells and sounds that will remain associated with that concept for a considerable time. Higher ability students will have formed a semantic representation of the concepts in a topic and are less likely to be led into misunderstandings and confusion by the particular wording or context in which a question is set than the lower ability students who have failed to convert their context-specific learning episode into a lasting semantic representation.

Phase 1 is Reading the question. Logically, reading occurs first and the other processes involved in answering a question either follow or emerge during the reading phase. Reading a question consists of constructing a representation of a *task* which is not only the *question* being asked, but also an *intention* about how to respond. Many misunderstandings and errors in answering exam questions occur during the reading phase. A question may be intended to assess students' understanding of subject-specific terminology, in which case reading failure may be valid. Otherwise we want to ensure that the question conveys to the student exactly what it is that we want them to do. The question's only function is to convey that task, and ideally it should be transparently easy to interpret. If we make the questions at all difficult to understand then we are constructing a reading test, not an achievement test. When students cannot understand the question they are prevented from showing us whether or not they can do the things we want to see. If different students are addressing different tasks, because some of them misunderstood the question, then we cannot assess them all fairly.

Reading a question in an exam differs from normal reading in several ways. There are competing demands, such as monitoring time during the exam and controlling anxiety, which reduce the available mental resources for processing text. These ‘stress’ effects can cause errors that seem unlikely in normal reading.

Phase 2 is the ‘Thinking’ phase which emerges from Phase 1. During the reading phase, concepts are activated in the student’s mind, and these in turn activate related concepts in a kind of cascade. Anderson (1983) calls this ‘spreading activation’. This does not mean that the student is conscious of the concepts being activated; indeed most of them will remain below the threshold at which the student becomes conscious of them. A few will seem particularly relevant to the task, because they match the specific features of the question; they will receive more activation, and rise into consciousness. These concepts will generate an idea of an answer to the task, and the student’s conscious mind will then be able to start planning how to write their answer.

All of these processes are fast and automatic. In normal life irrelevant concepts are activated only weakly and are quickly suppressed as activation is concentrated on the relevant ones. However, when students are reading exam questions they are under stress and are therefore more easily distracted by irrelevant ideas.

Phase 3 is the Writing phase. In this phase the student has to communicate their answer to the marker as a written response. The result of the previous phases however is just an idea of a response and is sometimes in the form of a multi-modal representation. Students must turn their multi-modal learning into a string of words, so that we can assess their understanding.

We would argue that most examination questions ask the student, in effect, to summarise a small part of their learning of the subject. This is quite obviously true for questions in Bloom’s Knowledge and Comprehension categories. To avoid encouraging students simply to memorise answers many other questions are set in contexts, or otherwise turned into Application category questions, but they still amount to an applied summary of learning. The same can be said of some questions that appear to demand Analysis. This process of writing a summary of one’s understanding has been shown to be a skill that improves with age (Brown and Day, 1982) and students at age 16 find it a difficult task.

We have used the Phases of the Model of question answering in order to develop a tool for analysing questions, which we call Outcome Space Control for Assessment theory (OSCA theory), described in the Section 5 below. The OSCA theory is also based on the idea of Outcome Space developed by Marton and Saljo (1976). This tool enables us to analyse exam questions alongside their mark schemes and identify those questions which are likely to cause invalid errors, that is questions which cause students to go wrong for the wrong reasons. Examples of questions that we think would cause problems at each phase are given in the Section 6.

4 Analytical tools

The methodological approach underpinning this study is grounded theory (Glaser and Strauss, 1967). Our Outcome Space Control for Assessment (OSCA) theory was developed using grounded theory, which is an explorative research technique, involving the researcher following the data. This contrasts with the hypothetico-deductive model, in which the researcher outlines a research hypothesis and each study sticks to that original topic. In grounded theory, the researcher sticks closely to the data, but is free to pursue the themes that it raises.

We began with the conclusions from earlier research (eg Pollitt et al, 1985; Pollitt & Ahmed, 2000; Ahmed & Pollitt, 2001; Pollitt, Walker and McAlpine, 2005), all of which involved sampling from national examination questions in Scotland and England and coding the kinds of issues seen at a surface level (first level codes), then at a deeper level (second level codes); in the case of the Pollitt & Ahmed papers, experimental studies and interviews were also used to generate data. From this, the main underlying themes were drawn out to form theories of how students process examination questions, and how examiners evaluate their responses.

4.1 Structure, questions and mark schemes

One other strand of earlier research that we utilised came from an experimental study funded by QCA into the effects of varying the level of ‘structure’ in exam questions (Pollitt et al, 1998). The findings were surprisingly complex, as increasing the amount of structure seemed to reduce some aspects of the demands in the questions while increasing others. Making questions more structured reduced the requirement for students to create their own structure for tackling the question or for expressing their response, but it also forced them to carry out the task in the particular way the examiners chose – structure reduces freedom as well as structural demand. Reducing structure made it easier for students to show what they knew, even if this was not what the examiners expected or wanted to see.

One clear conclusion, however, was that changing the amount of structuring altered the nature of the required cognitive processes. Since that research we have found it useful to distinguish three levels of structure, and to describe them in terms of the extent by which they **constrain** the student’s response. Very Constrained questions (VC), of which the extreme type is multiple choice, require a response in a more or less completely defined format; we include in the VC category many other questions where the answer is just a number, a word or a short phrase. With these the student has little or no freedom in how to answer the question, and the marker is essentially looking for a precise answer. Students are differentiated by *whether or not* they can do a task of this level of difficulty: assessment uses the *difficulty*, or ‘high jump’ strategy (Pollitt, 1991).

At the other extreme are Un-Constrained questions (UC), typified by the essay, in which there is little restriction on what the student may do to try to answer the question. Examiners expect to see certain concepts and content in good answers but generally evaluate the response in terms of the quality of the response. Differentiation is in terms of *how well* the students perform, using the *quality*, or ‘ice dance’ strategy.

Between these two, relatively pure, strategies we identify a category of Semi-Constrained (SC) questions, in which students are given some but not all the necessary structure for their response. In the GCSE exams we looked at in this study most questions belong to the SC category, with written responses ranging from a phrase to several sentences, or instructions to draw or modify a diagram. These answers cannot be judged *simply* in terms of being right or wrong nor *merely* in terms of the quality of the response, but need a combination of the two strategies. Both difficulty and quality are relevant in the marker’s decision about how many marks to award, making it more complex than either of the two simple strategies.

In terms of demands, the principal issue is that giving a question more structure reduces the demand on the student to create their own structure for responding to it. A response needs structure: if the question doesn’t provide it the student must. It is part of the examiners’ task to decide how important for their subject it is that students should demonstrate the abilities involved in generating response structure. Manipulating constraint or structure does not, however, only affect this demand, since it also changes the nature of the thinking processes required to complete the task. Giving more structure constrains the student's thinking, restricting the range of acceptable answers – what we call the *outcome space*.

4.2 Outcome space theory

Marton & Saljo (1976) introduced the concept of outcome space, in a study of how students approached the task of reading an academic article. They classified the range of responses given to questions about the article in terms of *deep* and *surface* approaches to reading the article, and described the different kinds of response in terms of how the students understood it. Given that learning is an idiosyncratic process in which the new information interacts with the student's existing knowledge they argued that we must expect different students to respond in different ways to questions about it. Even within a group of students of similar ability, who understand the new text equally well, there will be qualitatively different responses to a question. In scoring students' responses to UC questions we must, therefore, be prepared to consider at least two dimensions in the outcome space – one quantitative dimension representing the degree of success in dealing with the question, and at least one more qualitative dimension describing the ways in which the students attempted to respond.

We have extended this concept to SC and VC questions, particularly in order to understand what is required of a good mark scheme. A question's Outcome Space is the set of all responses to it, both actual and potential. The point is that, using the mark scheme they are given, a marker has to be able to evaluate every part of that space, every likely answer.

The set of all responses to a question can be separated into six subsets:

Good 1	The subset of	'Good' answers	Observed	Expected
Poor 1	The subset of	'Poor' answers	Observed	Expected
Good 2	The subset of	'Good' answers	**not** Observed	Expected
Poor 2	The subset of	'Poor' answers	**not** Observed	Expected
Good 3	The subset of	'Good' answers	Observed	**not** Expected
Poor 3	The subset of	'Poor' answers	Observed	**not** Expected

where 'Good' includes responses judged either correct or of high quality, and 'Poor' includes both wrong and low quality responses.

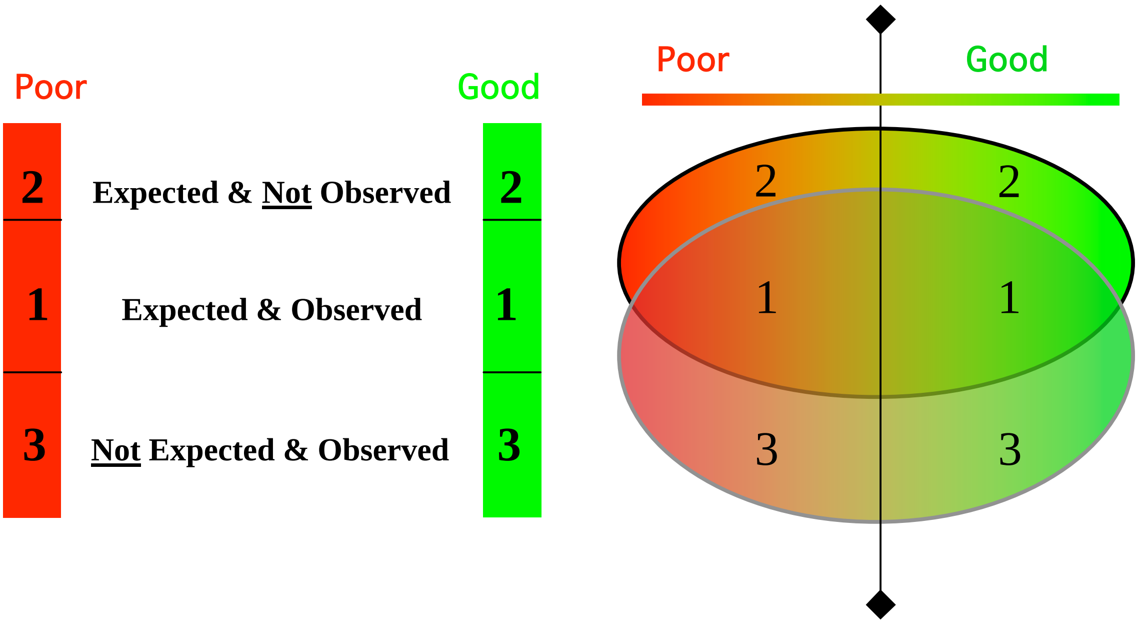


Figure 1. Outcome Space and its sub-spaces

We want the overlap between the intended and observed outcome spaces (areas **Good 1** and **Poor 1**) to be as large as possible, and the rest to be small. In particular area **Good 3** should be small as we don't want to see many right answers that have not been anticipated, while a large area **Poor 3** would mean that students are not behaving as expected.

4.3 Mark scheme classification

Pollitt, Walker and McAlpine (2005) developed a classification scheme for both questions and mark schemes, with the intention of optimising the fit of marking procedure to question type. This covered all of the general academic SQA examinations set in 2004 at four levels: Intermediate 1, Intermediate 2, Higher, and Advanced Higher. A modified version of that scheme was used to provide a systematic starting point for this project. An example of a completed spreadsheet from this is shown in the Appendices.

For each (part of a) question the command word/words used was/were noted, as a first indication of the kind of cognitive processes that ought to be involved in answering the question. The type of response – numerical; multiple choice; verbal, ranging in length and complexity from a word to extended writing; etc – was also recorded. The number of marks available for each question was noted, as well as any subdivision of it, such as where 4 marks were to awarded as '2x2' meaning that two similar responses were required for 2 marks each, or '2+2' where two separate parts were to be marked out of 2 each with its own mark scheme. Finally, comments were written to record any impression of problems with the question, the mark scheme, or the relationship between the two.

4.4 The outcome space generator

The technique of Outcome Space Generation, as described in Section 5, was applied to a sample of the questions. It would have been too time-consuming to apply this to every question, and unprofitable if similar issues arose repeatedly, and questions were therefore chosen whenever anything seemed particularly salient in the initial categorisation. The intention here was to generate, systematically, the range of all possible responses or types of response that a large group of pupils might reasonably be expected to produce in exam conditions.

The two examples in Section 6 will show, in considerable detail, how the technique was used.

As researchers, we are not examiners or teachers of each subject we study, and so may not accurately predict what most students are likely to respond to a question. But this also helps us avoid one constant problem that examiners face in preparing questions – they know, from their experience, what the examiner intends, and cannot read a question in the way that a 15 year old non-expert will read it while under exam conditions.

5 Method

5.1 Grounded theory

Grounded theory involves the following methodological features (Charmaz, 2006; Dey, 1999):

- *Theoretical sensitivity* - human knowledge is a construct, so rather than striving for objectivity, the researcher strives to uncover and account for her position with respect to the research topic.
- *Naïve wonderment* – the researcher should ideally have no knowledge of the topic, instead approaching it with no preconceptions about resulting theory.
- *Constant comparative analysis* – texts and the codes generated from them are contrasted to uncover different views of the same issue.
- *Coding* – the researcher codes findings conceptually, looking for the underlying meanings and the relations between them. Coding is carried out in layers, or 'tiers', with the first and second tier codes being largely atheoretical (Piantanida, Tananis & Grubs, 2004).
- *Researcher memoing* – codes themselves do not generate theory. Memoing is designed to assist the researcher generate and think through theoretical interpretations.
- *Saturation* is a feature of the method – the researcher stops sampling when no new information is being discovered.
- *Portraying the theory* – the persuasiveness of the final theory is critical to the evaluation of this approach. Positivist techniques of evaluation, such as verifiability, reliability, validity, and generalisability do not apply. Instead, factors such as rigour, ethics, integrity, verity, utility, vitality and ethics are central (Piantanida, Tananis & Grubs, 2004).

5.2 Data capture and analysis

Around 2,000 questions from GCSE question papers were sampled in total (Table 1). In this report we use the word *question* to mean one item intended to elicit a response from candidates, rather than to refer to the whole of a large possibly structured question. Thus "*Question 3b (ii)*" would be one question in this sense while the whole of "*Question 3*" might consist of seven or eight questions.

Board	Business Studies	Design & Technology	Geography	Total
AQA	88	96	86	270
CCEA	133	104	164	401
Edexcel	123	52	132	307
OCR	165	268	187	620
WJEC	101		214	315
Total	610	520	783	1913

Table 1 GCSE questions analysed[§]

Codes generated by previous research and forming the OSCA theory were used in the coding of the questions from the outset (see Figure 2). In the first analysis, each question was coded in a single line

[§] Approximately 150 other questions have been partially processed.

in a spreadsheet, recording the command words used, the nature of response expected, the kind of mark scheme used and how the marks were awarded.

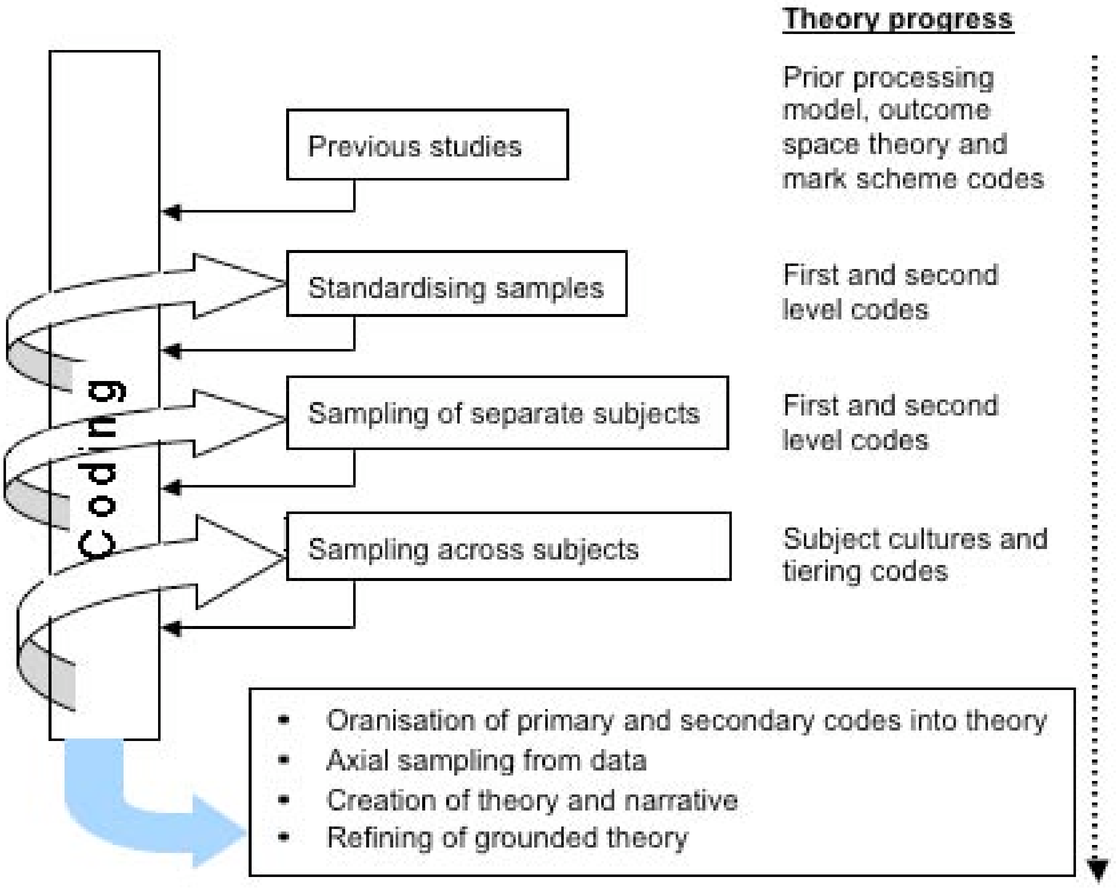

Theory progress
Coding
Previous studies
Prior processing model, outcome space theory and mark scheme codes
Standardising samples
First and second level codes
Sampling of separate subjects
First and second level codes
Sampling across subjects
Subject cultures and tiering codes
Oranisation of primary and secondary codes into theory
Axial sampling from data
Creation of theory and narrative
Refining of grounded theory

Figure 2. Data capture and analysis process

Three question papers were used to train and standardise the four researchers who conducted the coding:

Geography –	OCR Foundation Tier, Paper 1, 2006
Business Studies –	WJEC Foundation Tier, Paper 1, 2006
Design & Technology –	OCR Higher Tier, Paper 2, 2006

Frequent checks were made that raters agreed on the coding of randomly chosen questions.

Line by line coding was conducted for the standardising samples and for the early separate subject sampling. Beyond these samples, the focus was not so detailed because themes were repeating in the data. Some new codes were generated during the standardising process, but most codes were either already part of the outcome space generator repertoire, or were generated during the process of sampling separate subjects, or sampling across subjects. Once first and second level codes were generated, the narrative themes were drawn out and further (axial) sampling was conducted to flesh out those codes.

As the research progressed, some of the codes reached saturation and the n+1 rule was utilised, in which no further data is sampled if no new information is added by sampling an additional case (Piantanida, Tananis, & Grubs, 2004). Charmaz's (2006, p113) suggested questions were used to assess whether saturation had been reached:

- Which comparisons do you make between data within and between categories?
- What sense do you make of these comparisons?

- Where do they lead you?
- How do your comparisons illuminate your theoretical categories?
- In what other directions, if any, do they take you?
- What new conceptual relationships, if any, might you see?

Memos were used throughout the analysis process to note the relations between the codes being generated and to begin to formulate deeper conceptualisations of the data and codes.

5.3 Ethics

Ethical approval was gained through the School of Education at the University of Bristol.

6 Results

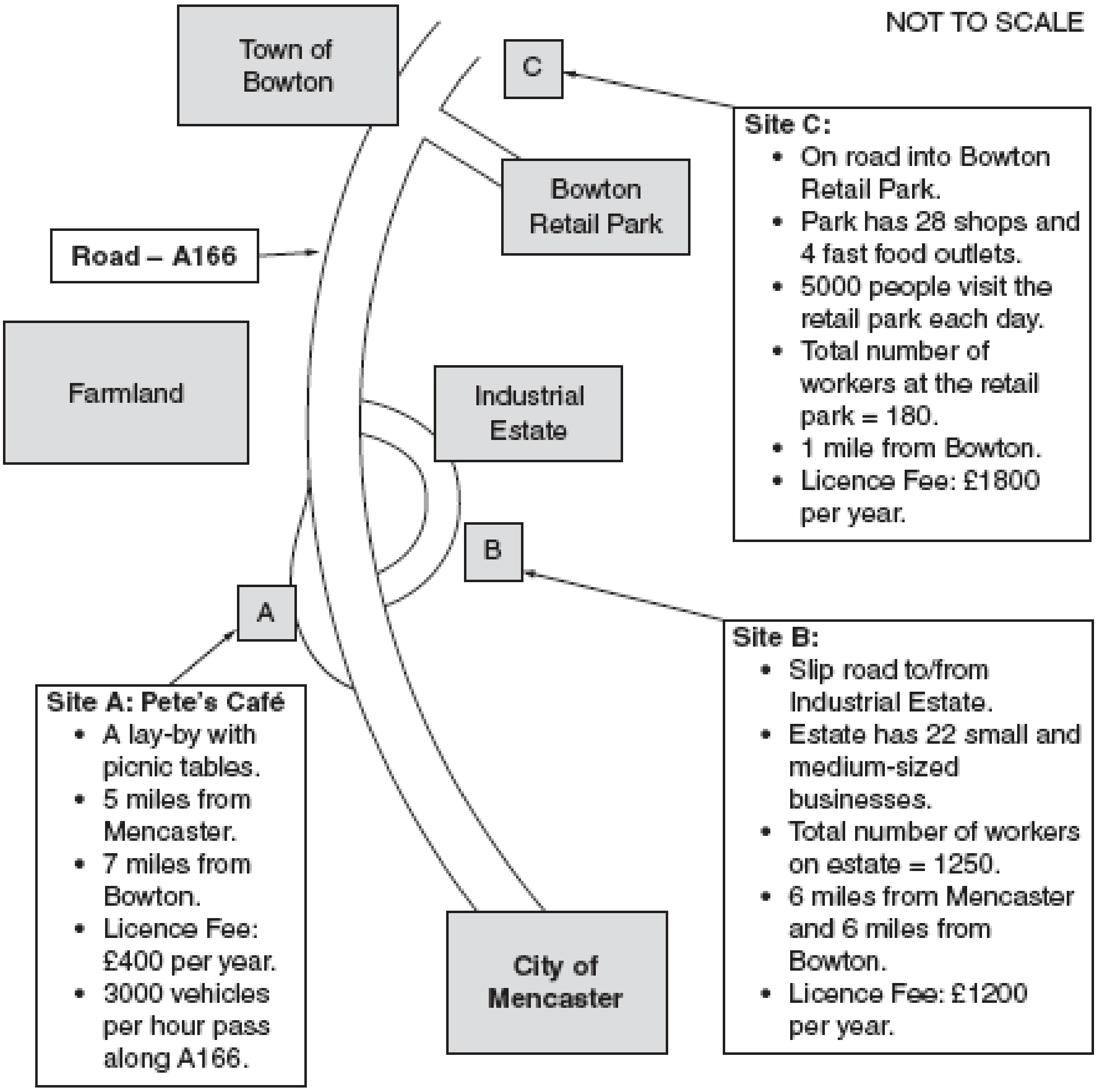
To illustrate our methods we begin this section with two examples of the analyses carried out, showing two questions that have been studied in detail, guided by our OSCA theory. In each case we show the question in its entirety and the mark scheme for the relevant part(s), the initial spreadsheet coding of the question and mark scheme, and the qualitative analysis that followed when we worked through the outcome space generator. The first example, from Business Studies, is shown in particularly full detail.

6.1 Example 1: *Business Studies, OCR FT P1 2006: Question 2(a) (iii)*

1 Peter Miller started a business called 'Pete's Café' in 2003. He sells drinks and snacks from a trailer parked by the roadside. Fig. 1 below gives information about:

- each of the sites he could have used for his business;
- the licence fee he would have to pay the local Bowton Council for using each site;
- the amount of traffic that passes by;
- business activity in the area.

Fig. 1 – Map of Moorshire County between Mencaster and Bowton.



Site C:
- On road into Bowton Retail Park.
- Park has 28 shops and 4 fast food outlets.
- 5000 people visit the retail park each day.
- Total number of workers at the retail park = 180.
- 1 mile from Bowton.
- Licence Fee: £1800 per year.

Site B:
- Slip road to/from Industrial Estate.
- Estate has 22 small and medium-sized businesses.
- Total number of workers on estate = 1250.
- 6 miles from Mencaster and 6 miles from Bowton.
- Licence Fee: £1200 per year.

Site A: Pete's Café
- A lay-by with picnic tables.
- 5 miles from Mencaster.
- 7 miles from Bowton.
- Licence Fee: £400 per year.
- 3000 vehicles per hour pass along A166.

1 (a) (i), (ii), (iii) **…**

1 (b) **…**

1 (c) **…**

1 (d) **…**

2 (a) (i) In the financial year 2004, Peter sold 25 000 items at an average price of £4.
Calculate the total revenue for Peter's business. Show your working.
....
....……………………………………..................[2]

(ii) It cost Peter £60 000 to buy the goods he sold. Using this information and your answer to question **(i)** above, calculate the profit that Peter made during the financial year 2004. Show your working.
....
....……………………………………..................[2]

(iii) State and explain **three** ways in which Peter could have increased his sales at Site A.

Way One ....
Explanation ....
...................................………………...

Way Two ....
Explanation ....
...................……………...

Way Three ....
Explanation ....
……………...…..................................[6]

2 (a) (i) Target: Ability to apply numerical skills to a business context.

Two marks for the correct answer, one mark for an appropriate method where the answer given is incorrect.

- £4 x 25,000 (1) = £100,000 (2).

Max: 2 marks

(ii) Target: Ability to apply knowledge of a Trading Account.

Two marks for the correct answer, one mark for an appropriate method where the answer given is incorrect. NB "ecf" rule to apply to total revenue value used.

- £100,000 - £60,000 (1) = £40,000 (2).

Max: 2 marks

(iii) Target: Ability to apply knowledge of how to increase sales revenue.

One mark for each appropriate suggestion, one mark for each point of explanation. NB Candidates can only get 2 marks for advertising or any other single method. Increase sales/ customers = 0 marks.

- Advertise (1) – by putting signs out on the road (1) or putting leaflets in the industrial units (1) this would raise awareness of his business (1). NB Media for advertising should be appropriate and specific e.g. local newspaper or specialist magazine).
- Reduce his prices (1) - to make his prices cheaper than his competitors (1) so that customers buy from him instead (1). Also cheaper prices might persuade some people to buy more things (1) or buy things they had not intended to buy (1).
- Give promotional offers (1) - such as buy one get one free (1)
- Provide a good service (1) like being friendly (1) extend opening hours (1) phone ordering (1) so people can collect a takeaway (1) which could be used by the industrial site workers (1).
- Increase the range of products (1) e.g. all day breakfasts (1)
- Packaging (1) – some description of the premises (e.g. sun shades, bright colour painting of trailer) or packaging used for the products e.g. provides pot mugs rather polystyrene cups.

Max: 3 x 2 = 6 marks

level	*FT*
paper	*P1*
year	*06*

Business Studies OCR

QP item number	question word	response type	marks	marking type	Notes
2a i	calculate	1.2 –Calculation	2	1.8 - Exhaustive list + rule	1 for arith
2a ii	calculate	1.2 –Calculation	2	1.8 - Exhaustive list + rule	1 for arith; literally, qn suggests he sold **everything** he bought - no wastage
2a iii	explain	2.2 – Short	3x2	2.3 - Points - examples	State and explain how P could have increased sales - 'get more customers' scores 0. So 'get more Cs by making promotional offers' scores only 1? So 'get more Cs by increasing range of products' scores only 1? In these cases the "explanations" for the 2nd mark are e.g.s Very wide interpretation of "Packaging", so paint the trailer to make it attractive, improve facilities at trailer to give better seats, buy new coffee mugs that people will like, to make the experience better - all scores just 2/6? Or do these at some point become "Provide a good service"? MS is about 'revenue', not number of 'sales'

Generating the outcome space

Before Phase 1:

We are looking at Question 2. Question 1 extended over two and a bit pages, and dealt with: the relative advantages of alternative sites for the café, what services the local council provide, Pete’s business objectives in the first year of running his café, the three sectors of an economy, and how employment levels in the secondary and tertiary sectors have changed. [Yes: all of these issues were in Question 1!] Students have read (we assume) all the information about Sites A, B and C from the map, and all the description of economic changes in Question 1 (d) and (e).

A student starting Question 2 has thought through these issues, extending from the very narrow and applied issue of the best site for the café up to the very broad and theoretical issue of trends in the overall economy. With Question 2 (a) they are brought back down to the most detailed level.

Phase 1 Reading the question

Parts (i) and (ii) are about money: calculating revenue and profit. Since language use is normally *coherent*, with a logical and predictable flow of ideas and functions, students will begin to read part (iii) with a presumption that it will be about money too. And, being familiar with the normal schema of examination papers, they might be looking for some issue that extends the concepts of revenue and profit, perhaps to consider their impact on Pete’s business. Apart from presumptions like this, there is no semantic content in the early part of the question.

"State and explain" will suggest to them that this is a structured question in which they have to write down something from the narrative and 'explain' it in some way; the next words "**three** ways" show that the response must be structured into three parts with two subparts each. At this point the candidate will:

- still be disposed to think about financial aspects of the business;
- maintaining in their heads a fairly complex structure for the answer;
- wondering what 'explain' means in this question;
- wondering what 'ways' might stand for.

They continue reading the sentence: "in which Peter" reinforces the notion that they must use the context of Peter and his business in their response. Then "could have" raises another ambiguity, concerning which kind of modality 'could' indicates here. Although they would certainly not be able to describe this ambiguity, and might not even be conscious of it, their minds will wonder whether this modal verb is epistemic or deontic: does it refer to things that it is possible Pete did in the past or does it refer to things he was permitted to do but did not do? Are they being asked to guess at Pete's history or to advise him after the event about what he could have done better? Readers are rarely aware of modal ambiguities like this, but their interpretation of the task they are being set will be biased by the way they unconsciously choose to process it.

They then read: " increased his sales" which, in the existing financial context from parts (i) and (ii) is likely to encourage them still more to think about money. At this point the competent readers will have constructed a complete mental model of a task that they might expect to be asked in this examination. There is a fairly complex specification for the parts of the task, a clear reference to the context, and a focus on increasing sales. Apart from resolving a few ambiguous points, the model is complete.

But the question sentence is not. If they are still reading carefully, they will now read "at Site A". If they pay enough attention to it, they will add it onto their mental model as a further element of specification of the task, but is very likely that many will not give it the attention it deserves. Under conditions of stress, and exams are stressful, many people tend to 'close prematurely', fastening their attention on the mental model as soon as they feel it is detailed enough to make sense. At the very least then, they miss some information that would have helped them address the task in the way the examiners intended them to: at worst they go on to answer the wrong question.

Even then the question is not complete. They will notice the layout of the paper and the spaces available for them to write their response. The layout here certainly reinforces – or corrects – the pupils' idea of the answer structure. Further, most will notice that 6 marks are available for their **three** ways and explanations, and they will assume something about how they are to be awarded; in this case the most likely inference is that there will be one mark for each 'way' and one more for each 'explanation'. [The mark scheme shows that they would be right if this is the inference they make.]

Phase 2 Thinking

Between the general context of 'Pete's Pantry', the immediate context of parts (i) and (ii), and the question wording and layout, there is plenty here to provoke the student's mind. We would hope (for the pupil's sake) that all the ideas that are prompted but are seriously irrelevant do get suppressed, but there is still plenty that they will continue to consider. They will search their memory (figuratively speaking, since almost all of this mental activity is automatic and so subconscious) for anything that might be strongly linked to the key concepts they have in mind – anything to do with 'sales', 'increasing sales', 'marketing', 'marketing strategies', etc, as well as anything they have met and considered before in class or in their textbooks or other reading. If, by chance, they know of a café anything like 'Pete's' they will probably search their experiences of being at it or simply passing it on the road.

Pupils with good knowledge of business principles will rely more on their formal learning and experiences of strategies a business might use to increase sales; they will also interpret the phrase 'increase sales' correctly as referring to increasing sales *revenue*. Others may rely more on their real world knowledge – it is one of the main problems for Business Studies that real world knowledge may sometimes give an unfair advantage to some students. Weaker student will fail to identify *sales* with *sales revenue*, and may think that increasing customer numbers alone amounts to "increasing sales".

Some may wonder if the word "ways" has a technical meaning in Business Studies – after all, *sector*, *factor*, *loss* and many others do. But somehow or other some ideas of 'ways' that might win marks will begin to rise into their conscious minds.

Phase 3-Writing

Turning the ideas of possible ways into coherent sentences is far from trivial. Writing is never easy for most pupils, and especially so when there are several constraints to consider. There are two particularly troubling words here: *explain* and *ways*. Suppose a pupil thinks "Advertising is often the answer they are looking for. Yeah, it would fit here". They would then write 'advertising' in the space: "Way One'. Now, what about 'Explanation'? What could it mean to "explain" advertising? Is it: "advertising is a way to increase sales because it makes more people aware that they could get a good cup of coffee at Pete's."

Or is it: "advertising is a way to increase sales because it makes more people aware that they could get a good cup of coffee at Pete's, and this would increase his sales."

Or is it: "advertising is a way to increase sales because it makes more people aware that they could get a good cup of coffee at Pete's, and this would increase the amount of money he gets from sales."

Or does "explain" mean "advertising is a way of spending some of your business money to inform and attract more people to become customers."

In fact, if we look at the mark scheme, it seems that "explain" here means something like *give an example*. The extra mark can be gained by writing: "advertising"; "by putting signs out on the road". This does suggest that there is no clear understanding amongst the examiners as to what "explain" *really* means, and it is very likely that the same confusion will be widespread amongst the pupils. If most of them are unsure what "explain" could mean in this kind of question, then there is likely to be a very wide range of kinds of answer given here, which would mean that the examiners have rather lost control of what the students' minds are doing.

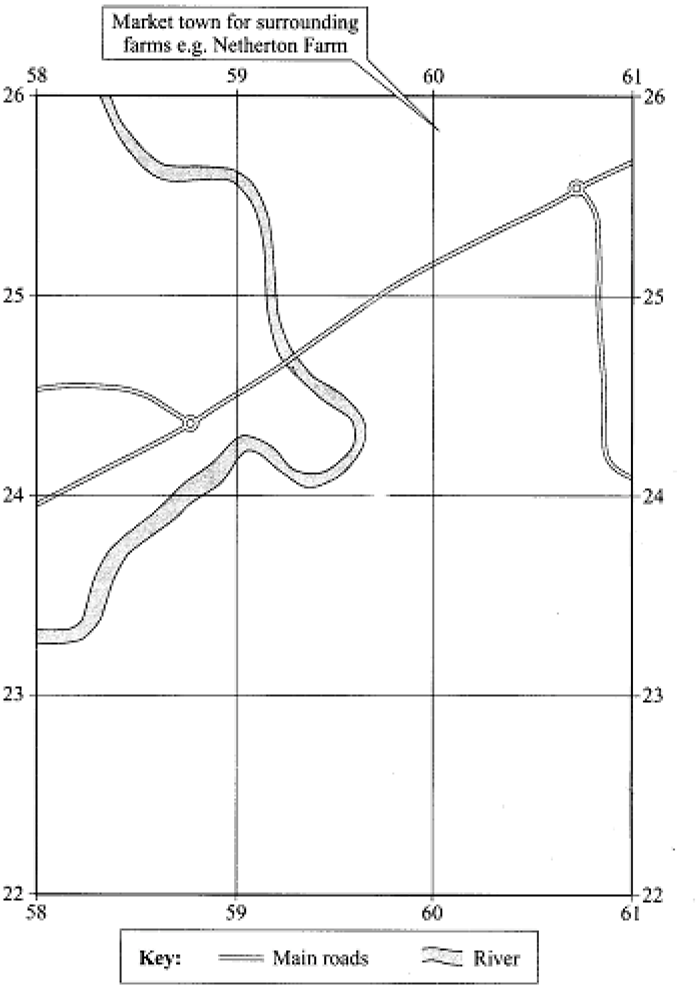
The other problem will emerge into the pupil's consciousness when they come to writing down a second idea in "Way Two". Suppose a pupil has thought of "putting signs on the road" and "advertising on local radio": how are they to decide that these are just one idea rather than two? Sophisticated abstract-thinking experts see these as two forms of the same strategy called advertising, but GCSE pupils are not experts and are much less likely to think abstractly. In any case, the question asks for "ways" which does not seem to be a technical term like 'strategies'. The mark scheme uses another term "packaging" in what seems an even broader concept, since it includes 'description of the premises', 'painting the trailer', using 'pot mugs' rather than 'polystyrene cups'. This is a very abstract concept of 'packaging' and one that most pupils are likely to find difficult.

6.2 Example 2: *Geography AQA HT P1 2006 Question 2*

Study **Figure 2**, a sketch map of Ross-on-Wye.

It is drawn at the same scale as **Figure 1**, the Ordnance Survey map extract.

Figure 2



2a Using Figure 1, shade and label on Figure 2 the built up area of Ross-on-Wye. (3)

2b Add detailed labels to Figure 2 to explain the location of Ross-on-Wye.

An example has been done for you. (4)

Mark scheme

2b Avoids flooding, avoids steepest land, close to bridging point, water supply, local resources, route focus etc. Defensive site. Fertile soils. Transportation.

Detail as in e.g. given = 2 marks i.e. dp

Brief point = 1 mark. Max for simple points = 3 marks. 4 marks

Initial analysis

level	*HT*
year	*06*

Geography

QP item number	question word	response type	marks	marking type	Notes
2a	label	3-Diagram	3	2.6 – Points /examples + scoring rules	
2b	label	3-Diagram	4x1	2.6 – Points /examples + scoring rules	will they guess they have to add **4** labels? *MS unclear* – what is a ‘simple point’?

Generating the outcome space

Before Phase 1:

We are looking at Question 2, and particularly at 2(b).

In Question 1 students were introduced to the idea of the settlement pattern of Ross-on-Wye, which will be relevant for Question 2.

In the preamble to Question 2(a) they are introduced to “**Figure 2**, a sketch map of Ross-on-Wye”, and asked to compare it to “ **Figure 1**, the Ordnance Survey map extract”. This is the third time the lengthy double noun phrase ‘Ordnance Survey Map Extract’ has been used. We now have four noun phrases all explicitly identified with each other – **Figure 1**, **Figure 2**, the Ordnance Survey map extract, and a sketch map of Ross-on-Wye. This all amounts to a fairly heavy burden on the student’s working memory.

They are then asked to shade and label “the built-up area” of Ross-on-Wye. This probably involves relating “the built-up area of Ross-on-Wye” from this question to “ the settlement pattern” from Question 1(c).

A student starting Question 2(b) will thus be primed to think in terms of settlement and other land use in the area covered by the map of the region around Ross-on-Wye.

Phase 1 Reading the question

The students have just completed a task that involved drawing on **Figure 2**.They now read “Add detailed labels to **Figure 2**”. It will be easy to understand the basic instruction of adding something to **Figure 2**, since that is what they have just finished doing in Question 2(a), but the object to be added – “detailed labels” – is not so straightforward. Maps do not normally have “labels”, even sketch maps, and they will need to read on to find out what this means here.

They next read “to explain the location of Ross-on-Wye”. The concept of the location of a town should be familiar to them, as it figures significantly in the syllabus. Once again, though, the rest of the phrase is problematic: labels on diagrams usually identify elements or describe some property of them, rather than ‘explain’ something. What is meant here by ‘labels explaining the location’?

Rather than puzzling about the meaning at this stage the student will continue to read, finding the second sentence: ”An example has been done for you”. Looking at the figure they find a label that says: “Market town for surrounding farms e.g. Netherton Farm”. The students should realise that they have to produce more labels like this one – but it is still not clear what “like this one” means. To resolve that, they must work out how the label ‘explains’ the location or Ross-on-Wye.

‘Explain’ has many meanings as a command word, and experience of exam answers shows that many GCSE children (and perhaps their teachers and examiners?) are not clear what the essential difference is between a description and an explanation. In this case good students should not have a problem, since the task should be familiar: it is to give a historical account of why the town is where it is, or why a town grew up exactly there. Students who are less competent or confident, however, may struggle with this. The reference to history is only implicit; it is part of the meaning of “explain the location” as a geographical instruction that it requires a historical explanation *rather than* a theoretical, political, or other kind of explanation.

Phase 2 Thinking

A pupil who is unsure about the meaning of ‘explain the location’ may be puzzled by the example. The ‘label’ says “Market town for surrounding farms”, yet it points to a place outside Ross-on-Wye, separated from it by a steep, narrow valley. The label doesn’t in fact point to the “Market town” but to an example of the “surrounding farms” which are mentioned at the end of the label text. To adults, and probably to most pupils in normal circumstances, this will not cause lasting confusion, but it is possible that some anxious pupils, under the stress of a high stakes exam, will be bewildered.

Pupils who understand the task, and understand that the label points to *an example illustrating a reason* why Ross-on-Wye is located where it is will now ‘search’ their memory of geography for other reasons for the location of towns. The intention is that they should recognise examples of several common reasons, such as the presence of a river, a bridging point, or high flood-free land, and demonstrate that they understand why these features were historically significant in the origin of the town. Since they have just shaded in all of the built-up area of the town in Question 2(a), it is possible that some students may interpret the word ‘location’ to include the whole extent of the present town, missing the implicit reference to history. They might then, for example, look at a modern suburb north of the town and label it as ‘Next to the A40’, even though this road is a modern by-pass.

One of the standard ‘reasons’ for the location of a town is the existence of a suitable bridging point, and this is certainly true for Ross. But there is a twist to it here: the sketch map shows only one bridge, the modern one built for the by-pass. The bridge that explains the origin of Ross is shown on the proper OS map but not on the sketch map. Markers are not told whether a mark should be awarded for labelling the wrong bridge, or only for labelling the original crossing point. It seems that the markers have missed a chance to distinguish between average students who know a crossing point is a ‘reason’ for the location of a town and better students who can also identify and label the right spot for it.

Phase 3-Writing

Students who have ‘solved’ the problem by understanding the task and knowing what features to label are still faced with at least two puzzles in the final phase of producing their written response. First, the label they are given on the sketch map as an example is written outside the map itself and there is not very much other space outside the map for them to use for their own labels. They know that the task is to ‘explain’ the location of the town and they are used to writing quite a lot of words when they write an explanation; some of them will feel the lack of space as a conflict with the task demand. Should they put labels on the map? How much should they write? They are asked not only for labels but for “detailed labels” – what exactly are “detailed labels”?

This leads to the second puzzle: how many labels should they add to the figure? If the question doesn’t make this explicit it is customary to look to the number of marks available for guidance. In this case there are four marks, but there is nothing to help pupils decide between giving four labels and expecting one mark for each or giving just two with enough ‘detail’ to get a second mark on each. They might also look to the example for guidance. It contains three elements – the function ‘Market town’, the justification ‘surrounding farms’, and the example or position of the label ‘Netherton Farm’. That seems quite a lot for one mark so perhaps two labels will do? But there are quite a lot of features they could label, and no easy way to see two that are the most important ones, so the dilemma cannot be resolved. We, of course, can look at the mark scheme to see what was wanted, but the pupils were not able to do so.

This final dilemma flags for us a very general problem with the assessment strategy often used in GCSE. The task to “explain the location of Ross-on-Wye” is quite clearly a valid geography assessment task that can be completed well, moderately, poorly, or not at all by different pupils. Constraining the task in some way, such as by asking for "labels … to explain the location", may harm the task's ability to make the very discrimination we are looking for between different degrees of competence in geography. In this question it is possible for pupils to score the same mark in very different ways, some of which will be clearly better than others. It is possible to get full marks, for example, without mentioning the river, but this would surely not be a very good 'explanation'. The mark scheme is based on counting 'correct' points rather than on assessing quality.

The mark scheme gives credit for 'detail as in example given', so that neither markers nor students have any more guidance on what is meant by 'detail' and how to get that second mark for each example. If students give no detail, just 'simple points', they can score a maximum of three marks, i.e. they will only score three marks even if they give four simple labels. This is not indicated to the students in any way.

These examples show, in more detail than normal, how we used the "outcome space generator", a procedure for thinking through the students' thinking processes carefully, trying to spot any conceivable difficulties some of them might face, and judging whether these amount to valid or invalid sources of difficulty. For this study we found it an excellent way of identifying features of the questions and the mark schemes that were worthy of coding as data for later explanation and generalisation.

7 Procedure

As exam papers and mark schemes were analysed many issues arose. These were noted, and checked by looking for other, similar or contrasting, examples in the same or other papers. From lists of detailed observations, both in the initial spreadsheets and more informally, each member of the team suggested generalisations that seemed to describe several of the phenomena seen; through discussion and exchange of written notes, and by the application of such theoretical ideas as had already been developed, some tentative themes and principles were proposed, tested and refined.

Rather than present all of the stages of this process, we will present our findings and discussion structured according to these emergent generalisations, themes and theories. We will illustrate the points we raise with examples from the examination papers and mark schemes. After that, we will suggest some strategies that may be useful to examiners seeking to improve the quality of the examinations for which they are responsible.

8 Findings and discussion

In this section we catalogue some of the problems we found with questions and mark schemes in this study. Around fifty examples are given where we believe the validity of the exam was compromised in the way indicated – chosen from the 2000 or so questions we studied.

We begin by restating our governing conception of validity for exam questions and mark schemes:

An exam question 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.

Thus the business of a question writer is *mind control*: to use the wording and presentation of the task in order to ensure that every student's mind is trying to do the things we want them to show us they can do. The business of a mark scheme writer is to ensure that the credit given to each student accurately reflects how well they did it.

In this study we are looking for ways to improve the quality of the assessment in GCSE. We have therefore concentrated on looking for problems, examples where we think the questions will not have worked as well as intended. As a result, the tone of the following discussion will be predominantly critical, *even though it is apparent to us that the general quality of question writing in the 21st century is better than it was in the 1980s, when we began to study these issues*. Our criticisms are intended to be constructive.

The essence of good examining appears to be the congruence of several ideas:

the question writer's idea of the task they want students to carry out to show their learning;
the students' understanding of that task and what they are supposed to do about it;
the marker's idea of what should be given credit in answers.

In our study of these exams we found cases in which these three ideas seemed less than perfectly matched. Sometimes the question expressed the idea of the task poorly, and we think students, especially the less confident or more anxious ones, will have found it unreasonably difficult to understand what they were supposed to do. Section 8.1 therefore addresses *failure to control the students' thought processes*.

The second part of the conception of validity insists that the mark scheme should ensure that markers give proper reward for evidence of appropriate thinking. We found examples where the mark scheme did not give enough, or clear enough, help for a novice marker. Of course, the mark scheme is not the only resource available to a marker, and a weak mark scheme is not necessarily a serious problem, but we feel that there is considerable scope for improving the validity of examining in addressing *inadequate mark schemes*, the topic of Section 8.2.

On other occasions we felt that the mark scheme did not give credit in the way the question suggested it would, or in the way that we thought appropriate for the skills being sought. Even if the mark scheme itself were well written it might not suit the evidence students were giving. In Section 8.3 we address cases where exam validity may be threatened by *mismatch between the question and the mark scheme*.

Finally, in Section 8.4 we have collected a few *other issues* that don't fit into these main categories.

In the discussion of the examples we will sometimes refer to the six subspaces of the Outcome Space shown in Figure 1. The spaces called **Good 1**, **Good 2** and **Good 3** refer to the subsets of correct or acceptable responses, while **Poor 1**, **Poor 2** and **Poor 3** refer to wrong or unacceptable ones. The main concern is usually with **Good 3** and **Poor 3** – the subsets of unexpected responses – since these often show that the question did not function in the way the examiner intended.

8.1 Failure to control the students' thought processes

It has long been standard practice to consider the kinds of thinking that students would need to use to answer questions, at least since the development of Bloom's taxonomy of educational objectives in the 1950s (Bloom et al, 1956). The most popular approaches to this have been aimed at the 'command words', the verb phrase or verb that 'tells' the candidate what kind of response to give. But if we want to ensure that the students' minds are doing the things we want them to show us they can do, there is more that needs to be considered than this.

In this first section we are concerned with poorly written questions, poor in the sense that they do not induce the students' minds to do the kind of thinking intended, or at least that there seems a significant danger that they will fail to do so in the case of some candidates. We begin with cases where, quite simply, the 'wrong' skills may be used to answer the questions.

8.1.1 Guessability

If the students' minds are doing the things we want them to show us they can do, they should not be gaining many marks by guessing. Consider this example:

Example:

(iv) Cross out the **wrong** words in the sentences below.
Japan is an **MEDC / LEDC**.
The percentage of people employed in tertiary industry has
decreased / increased because these jobs pay **more / less** money. (3)

While multiple choice questions may be appropriate in GCSE exams it does not seem wise to use 2-option binary questions like these, since they allow luck to play too prominent a role in a pupil's score.

8.1.2 Testing English

A further issue with the 'Cross out the **wrong** words' questions is their dependence on grammatical and lexical skill. Consider this paragraph:

Example:

(d) The paragraph below describes the formation and features of a delta.
Choose the correct words from the box to complete the paragraph.

distributaries	deep	heaviest	shallow
erodes	tributaries	lightest	deposits

A delta forms where a river flows into a .... sea.
The river .... a lot of material, which builds up.
The .... material is dropped first.
Because there is so much material, the river splits into .... (4)

Despite the appearance of a choice of 4 from 8 options, grammatical knowledge alone turns these four items into binary choices: the pairs are, respectively, simple adjectives, verbs, superlative adjectives, and nouns. Candidates who wrote a grammatically wrong answer would be showing evidence of poor English rather than of poor geographical knowledge.

Taking this effect into account, there were 27 marks awarded in this particular paper on binary questions.

8.1.3 Use of real cases – understanding or memory?

There is a potentially serious effect of using real contexts or cases, as is common in Business Studies and Geography. The mental processes students use to answer them may be quite different if the case was salient enough for them to remember it, or for their teacher to have used it as an example in teaching. The intended process, perhaps of showing understanding or of applying knowledge to a practical example, may turn into simple recall.

In 2006 a Geography exam used the Boxing Day 2004 Indian Ocean tsunami as an example:

Example:

What caused the loss of life in the affected countries? (1 mark)

In simple cognitive terms it is very likely that many pupils will have remembered this event from just 17 months earlier, and will need no geographical knowledge to answer the question. Emotionally loaded events are remembered more than neutral ones, and recalling them brings back the emotion they experienced at the time. Besides the 'recall' rather than 'understand' effect, this raising of emotional level during an examination may interfere with some children's cognitive processes, reducing their level of performance.

8.1.4 Use/Misuse of emphasis

Key words in a question can be emphasised to prevent pupils missing or misunderstanding them in their reading: after all, we want the students' minds to be doing the things we want them to show us they can do, rather than to test their reading accuracy. But there is a tendency for advice like this to be turned into rules which are then followed without much thought; typically, in this case, there is a rule: 'always put numbers in bold'. Consider this example:

Example:

(c) In the period from July to October, Fruitizz had reached the Maturity and Saturation stages. Distinguish between these **two** stages. [4]

The bold "**two"** does not help here, and the word should not only not be in bold but should be omitted altogether. Sometimes the bold number is even less helpful:

Example:

(iii) Identify and explain **one** piece of legislation that the zoo will have to consider when promoting its services. [3]

The important phrase here is "when promoting its services", not the idea of 'one' piece of legislation. Highlighting **promoting its services** might have helped direct pupils **away** from irrelevant legislation concerning animal welfare or employment.

Bold could be used more positively than it usually is:

Example:

(i) Since 2005 many powerboats have stopped using Lake Windermere because of a new 10 mph speed limit. **Figure 11** shows some people who have opinions about this speed limit.

(Pictures of Powerboat owner, Visitors who walk and sail, Local shopkeepers,
Local petrol station owners, Local residents)

Using **Figure 11**, choose **one** group of people that is against the speed limit on Lake Windermere and say why. *(2 marks)*

The question is difficult to understand correctly as it contains a double negative: 'against' and 'limit'. Candidates have to identify those who are against limiting the speed i.e. those who are *for* powerboats using the Lake. It is very easy to miss a negative because the sentence will still make perfect sense without it. In this case interpretations such as 'people against speeding' will be common. Highlighting the word **against** would have helped here, or phrasing the question in the positive. It is worth stressing that there is much more to **negative** than just the word **not**.

Another example:

Example:

(d) Give **two** quality control checks that could be carried out during manufacture of the clothes hooks.

1 ____[1]

2 ____[1]

Here again it would have been more helpful to stress either **during** or **manufacture**, rather than ‘two’ – which is clearly indicated by the layout of the answer space.

Finally, there is one question where we can state with reasonable certainty how the pupils will have responded:

Example:
(ii) Describe, in detail, the shape of the valley along this cross-section. [3]

We met a similarly worded question several years ago (the earlier version did not contain “, in detail,”) and we manipulated it in an experiment by replacing ‘shape’ with ‘**SHAPE**’. The result was an increase in the success rate from 8.3% to 37.5%. In the recent question we would expect exactly the same to happen as in the older one – many pupils will fail to concentrate on the *shape* or, to put it another way, their minds will not be doing the things the examiners wanted them to show they could do, that is, describe a hanging valley. With the help this time of a diagram, the question seems to have been more successful, but the Examiners’ Report states that it was:

> *well answered, particularly by those who restricted their description to valley shape in part (ii) rather than land use.*

It seems that, once again, highlighting the word *shape* would have improved the validity of the assessment.

8.1.5 Other reading difficulties

The zoo question above illustrates another source of reading difficulty, which can also be seen in the next example:

Example:
(b) Discuss whether penetration pricing would be the most appropriate pricing strategy for Center Parcs to use for its new conference business customers. *(8 marks)*

The ‘Report on the Examination’ contains the following remark:

> “It was a concern that many candidates failed to read the question fully and disregarded the intended target market, conference business customers. Many spent too long explaining the meaning of price penetration …”

But this was inevitable, given the linguistic structure of the question. The task phrase “Discuss whether” is followed by the Subject-Verb-Complement structure “price penetration – would be – the most appropriate pricing strategy”: at this point the sentence seems ‘complete’, and anything that remains will tend to be treated as of lesser importance. If the market segment was meant to be crucial it should have been placed first, as in:

“Discuss whether, for its new conference business customers, penetration pricing would be the most appropriate pricing strategy for Center Parcs to use.”

Processing is made still more difficult here because the noun phrase “its new conference business customers” is complex and difficult to parse. Examiners should accept that it is their responsibility to convey clearly to the candidates what it is they are expected to do, rather than just express ‘concern that many candidates failed to read the question fully’.

8.1.6 Ambiguity

There is a more serious error that examiners commit occasionally, but where it is more difficult to blame them – writing an ambiguous question. For example:

Example:
(i) Photograph A shows a flood bank. Why was it built here? (1)

MS Point mark

- to prevent the houses from being flooded (1)

The mark scheme addresses the question “Why was it built?” with only a nod towards the word “here”. A more sophisticated geography question would concentrate on the *location* of the flood bank rather than its purpose, which is rather obvious. In this example it is not clear which of the two possible questions the examiners intended, as they do not give enough guidance in the mark scheme to deal with answers such as

‘to prevent flooding’ or ‘to protect the houses’.

We cannot tell which of these answers belongs to the expected or the unexpected categories of response: **Good 1** or **Good 3**. It looks as if the examiners did not notice the ambiguity, but we can be sure that some pupils did. That is the real practical problem with ambiguity – it often takes the *non*–expert mind to spot it.

If the ambiguity is noticed, it may be possible to handle it in the mark scheme:

Example:

(*e*) Suggest and evaluate ways in which Emma can expand her business. [7]

MS Level 1 [1-2] Suggests way(s) in which the business might expand but with no evaluation.
Level 2 [3-4] Suggests way(s) in which the business might expand with one - sided or unsophisticated evaluation.
Level 3 [5-7] Suggests way(s) in which the business might expand with well-balanced/sophisticated evaluation.
Suggestions might include:

- merger; (many *more points indicated*)

Some candidates may interpret this question as the method by which businesses expand and this must be fully credited. Others may look at it in terms of raising finance for expansion again this must be fully credited.

It is the last paragraph that is of interest. This is a good example of a mark scheme in that it recognises alternative interpretations of the task and advises markers how to deal with them, even though the very existence of these acceptable alternatives (**Good 3**) indicates that the question failed to control the students’ minds as originally intended.

A common source of ambiguity is the use of modal verbs like may, might, can, could, should or would, which are used in English to convey several senses of uncertainty or obligation. Consider this example:

Example:

(e) Explain how Becky and James might use technology in their business to maintain a competitive edge and to achieve high quality service for customers (see the case study lines 42–46). [6]

The problem here is with the modal verb ‘might’. Candidates already knew from the pre-release materials that Becky and James *do* use technology:

> “Over the years, Becky and James have made considerable use of technology in their business in order to maintain a competitive edge and to achieve high quality service for customers. This technology includes word processing, digital cameras, databases and mobile phones. Also, the increasing use of the Internet by many prospective clients means they must constantly review the emphasis of King & Khan’s marketing strategy. Despite this increase in the use of technology, Becky and James are still concerned about the need to revise their advertising methods to sell properties.”

Which kind of modal verb then is ‘might’? Is the question about *probability*, or suggesting what they are *already* doing with technology to achieve that aim? Or is it about *possibility*, or what they might do *in the future*?

8.1.7 Vague command words

When a command word is modified by an interrogative adverbial the result may be a poorly defined question. In the example below, ‘describe’ has been changed into ‘describe how’, and a modal verb ‘could’ is also used. It is then unclear what kind of answer is expected. What level of detail? Are technical terms needed? What kind of sample? The mark scheme is formal and general and more about the principles that *should* be followed than a description of how it *could* be done.

Example:

(c) The prototype device would need to be trialled before the product is manufactured in quantity.
Describe how trials of the prototype could be carried out.

MS: Prototype trials involve giving individuals the opportunity to try out the device,
and obtaining feedback either by means of questionnaire or interview [2]

8.1.8 Multiple command words

The command word is a direct instruction to the candidate to do something. It is therefore usual to have just one command word for each question, or to use one of a fairly familiar set of paired command words like 'state and explain' or 'compare and contrast'. Unusual combinations can cause trouble for both pupils and examiners:

Example:

(c) Resource exploitation can damage fragile environments.
Choose a case study of a fragile environment damaged by resource exploitation.
Chosen case study ....
Describe the causes of the damage to the environment
and
explain the effects on the environment and on the local people and explain what is being done to manage the problem. 7

The topic of the question is introduced clearly – it is about resource exploitation. Students then have to choose a case study. Then they must describe the causes of damage to the environment. Then they must explain various things. This 'explain' part of the question is intrinsically difficult to process as it asks for three different things in one sentence: effects on environment, and on local people, and what is being done.

The whole task is thus very complex for pupils, and it is likely that many will perform unevenly on the several aspects of it. The mark scheme is a three levels of response scale of the 'best fit' type but has little guidance on how to combine the various two parts to give a final mark.

8.1.9 'Explain'

There is one command word that seems to cause as much trouble as all of the others together – the word 'explain'. One of the reasons it is problematic is that it can mean quite different things in different questions.

Different meanings

Consider first what may happen if examiners ask a direct 'why':

Example:

(i) Why do workers, such as Jane Price, pay National Insurance Contributions? [1]

MS Any one from the following:
Statutory deduction from pay/must pay
Entitles Jane Price to state benefits
Other valid points

The phrase "such as Jane Price" is an unhelpful use of the context here; it encourages pupils to try to work the context into their answer when the question is actually assessed entirely free from context. But more serious is the qualitative range of the answers in the mark scheme. The two points listed are utterly different in nature, and correspond to two different kinds of explanation – the first says she pays because legally she has no choice, the second explains why we accept having no choice.

It should not surprise us, then, if candidates have difficulty interpreting 'explain' questions in the way the examiners intended:

Example:

(c) Identify and explain the type of management structure that exists in Cafedotcom. [3]

MS Flat [1]
This means the business will have a short chain of command and could lead to larger spans of control [2]
([1] + [2]) [3]

As the mark scheme shows, this 'explain' was supposed to mean something like 'define' or 'state the meaning of'. However, the Chief Examiner's Report notes:

> The type of management structure was usually correctly identified as flat although some candidates failed to explain that structure and instead described the roles of the management team and the line of authority.

We believe the candidates were unsure what '*explain* the type of management structure' meant.

Depth required

A further, and particularly difficult, problem with 'explain' relates to the depth of explanation required. In the 'Jane Price' example the single mark available was given for either a shallow or a deep explanation. Sometimes only one level of explanation is accepted and candidates may miss out by giving an 'explanation' that is either too deep *or* too shallow.

Example:

(ii) Explain how the change shown in **Fig. 6** improves the safety of the crusher. **[2]**
MS Both hands are used to operate the crusher/both buttons have to be pressed [2]

The 'model' answer in the mark scheme doesn't explain *why* using both hands makes the crusher safer: isn't it having to use both hands that makes it safer to use the crusher? That it can't be switched on with out full attention paying given to it?

Some further examples of this difficulty with 'explain' will appear later.

'Explain' a problem

We found examples of specific confusion where 'explain' is combined with certain nouns. One board regularly asks candidates to 'explain' a problem:

Example:

(e) Suggest and explain **one** problem faced by businesses which use an organisational structure like the one shown. [2]

MS Suggestion. [1]
Explanation. [1]
Answers might include:

- length of chain;
- messages may be lost/misunderstood;
- change may be resisted by those down the chain;
- motivation may be low for those at the bottom of chain;
- etc.

In this mark scheme it is hard to see whether the 'answers' a candidate 'might include' are suggestions or explanations. Suppose a pupil wrote:

> *The length of chain, because messages might get lost.*

Is that worth two marks? Or do they have to explain *why* they might get lost? Or do they have to explain *why* losing messages is bad for a business? It does look as if 'mentioning' a problem is often treated as equivalent to 'explaining' it.

'Explain' a reason

What does it mean to 'explain' a reason? This is logically problematic.

Example:

(i) Explain one reason why good advertising would be important to Cafedotcom. [2]

Does it mean that candidates should give a reason and then explain its importance to the business? But the mark scheme is:

MS Good advertising would be important because:

- Cafedotcom will want to create awareness
- Potential customers will want to know when it is opening [2]
- Increase sales

(1 × [2]) [2]

There is no explanation required, of why it is important to a business to 'create awareness' or tell people 'when it is opening', yet full marks is awarded for these answers. It seems that this 'explain' just means 'state'.

'Explain how'

Is 'explain how' different from 'explain why'? There are dialects in Britain where *how* is used almost as a synonym for *why*, especially in phrases like 'How not?' or 'How come?' A literal use of 'explain how' can run into problems:

Example:

(ii) Explain how the names of ingredients are arranged on the packaging of ready prepared pastry. (2 marks)

MS Written in order … descending order by weight
largest first
legal requirement (2 marks)

The first two answers in the mark scheme are indeed *descriptions* of *how* the names are arranged, but it is arguable whether a description amounts to an explanation here. The third answer – "legal requirement" – tells us *why* the names of ingredients are given. We presume that the legal requirement is for the names to be listed in this way but, again, it is arguable whether this is an explanation: one could still ask *why* the law requires this rather than another way. As for the mark scheme, does it demand the legal answer for 2 marks? Or even for just 1? We suspect that most candidates did not think carefully about the intended meaning of 'explain how' but simply gave the response that seemed appropriate to them on first reading the question. For most, that would probably be the 'describe how' version. The Report on the Examination records:

> Question 6 (b) (ii) Many candidates attempted this question but could not respond with the clarity required to gain two marks

which may amount to confirmation of our impression.

'Explain' a thing

Example:

> In the late 1990s Nicholas Hayek, a Swiss businessman, worked with the well-known luxury car manufacturer Daimler to finance and produce a vehicle which was sold under the Smart car brand. The low price, two-seater Smart car was Daimler's attempt to enter the small car market and to break out of the luxury market.
>
> *Adapted from www.bbc.co.uk 19 December 1999*

(a) Suggest and explain the market segment at which the Smart car was aimed. [2]

What does it mean to "explain … the market segment"? The mark scheme does not state a 'correct' answer, but allows any of several ways of segmenting the market; it seems that the point of the question was to find out if candidates can properly distinguish between the technical term ‘segment’ – referring to people – from words like ‘sector’ or ‘market’ which refer to goods. But it is still odd to ask them to “explain” it; ‘describe’ would have been better.

8.1.10 Effects of context

One of the reasons we chose to study these subjects is because they make extensive use of *context*; in fact, almost all of the questions in all three are set in particular business or geographical contexts or in the making of a particular product. Despite the very good arguments for doing this, it must be recognised that contexts can cause problems for both candidates and examiners (Ahmed & Pollitt, 2007). We have already noted the problems associated with using real contexts rather than made-up ones, but there are others.

Example:

6 Center Parcs believes that it is important to have a well-trained and motivated staff (see **page 9**). We are told that Center Parcs tries different ways to motivate employees.

(a) Describe **two** difficulties which Center Parcs might experience by having a large proportion of its employees working part-time. *(6 marks)*

This question is set within a large context, in that the whole paper is set on a single case study of Center Parcs, which was pre-released to centres. But there is also a 'micro-context' effect that will operate here. The first synoptic paragraph will focus candidates' minds on issues to do with training and motivation, but these are *only* relevant to question (b) and not to question (a). The mark scheme is:

MS Possible reasons include:

- lack of commitment to the organisation;
- communication problems;
- finding sufficient people in the area;
- difficult for teams if people working different hours.

If students' minds are directed into irrelevant areas they will *not* be doing the kinds of things we want them to show us they can do. Similarly:

Example:

7. Brian Gregg and Davy Packham once played professional football in the lower divisions of the football league. After they retired they decided to set up their own sports shop.

They have found an empty shop for their business but it is located fairly close to a large sports shop owned by a national chain.

Brian and Davy decide that they will go ahead with their plans but they realise that it will be important to advertise their business.

(a) Discuss the main arguments for and against Brian and Davy setting up as a partnership. [6]

Candidates will expect the lengthy text they have just read to be significant in answering the question. The mark scheme contains a list of fifteen 'arguments', but none of them are particularly relevant to "Brian and Davy" and several points seem quite irrelevant to them (such as 'continuity'). The final point in the pre-text concerns advertising; as in the previous example the jump from 'advertising' to 'partnership' is linguistically incoherent, and candidates will struggle to make a link.

We have observed this same problem in other GCSE exams, and on one occasion challenged the examiners to justify it. Their intention was to give pupils a simple and easy start to the whole structured question, by setting a question which is "fairly straight-forward textbook stuff". But our evidence showed that this strategy failed; candidates did not understand the examiners' plan and were indeed misled into trying to use the context in the first question.

Sometimes context tempts examiners to ask unanswerable questions:

Example:

8. Julian Smith is eager to set up his own business after gaining a degree. He believes that there are opportunities for him in his own city of Swansea to set up an office cleaning business.

(a) Suggest **three** reasons why Julian wants to set up his own business. [3]

Since we do not know Julian we have no way of knowing what motivates him. One wonders what kind of degree he gained.

8.1.11 Testing other skills

We have mentioned reading skill as one irrelevant skill in assessing learning in other subjects; there are others. Sometimes an answer may depend unreasonably on mathematical skills, as in this example:

Example:

(c) The above chart shows the total number of people in 2005 who went to see the films shown in one EPP cinema.

(i) Which was the most popular film? (1)

The graph showed the attendance at six films. To get the single mark candidates had to read the graph accurately six times and add up the six readings accurately. The sum was:

15,000 + 25,000 + 20,000 + 30,000 + 60,000 + 40,000 = 190,000

which is not a trivial sum for a Foundation Tier candidate to calculate.

Also, we have a concern about the number of separate resources that Geography candidates are sometimes required to manipulate simultaneously.

Example:

2 Study **Figure 2**, a sketch map of Ross-on-Wye.
It is drawn at the same scale as **Figure 1**, the Ordnance Survey map extract.

Figure 2

(next page)

(a) Using **Figure 1**, the Ordnance Survey map extract, complete the following on **Figure 2**:
Label the A40(T) *etc*

Example:

(a) Look at Figure 1a. It is a sketch map of Blandford Forum.
Also look at Photographs A, B and C in the Map and Photograph Booklet.

Figure 1a

(next page)

(i) Complete the sentences below by crossing out the wrong word.

etc

In these it is at least arguable that it is more difficult to make sense of the task than to complete it. In the next example the added resources certainly get in the way:

Example:

(a) Look at Figure 3a. Also look at Photograph D in the Map and Photograph Booklet.
These show tourist activities on a beach in Cuba.

(i) **Active** or **passive** is one way of classifying tourists.
Use the words **active** or **passive** to complete the labels on Figure 3a.

Sailing is
....

etc

In this case Figure 3a is a tracing of the principal features in Photograph D. The photograph itself adds nothing to what is already given, and is simply a confusing distraction.

8.2 Inadequate mark schemes

Questions are the obvious focus for improving examination papers, and question writing has received attention for several years. But however good the question, its contribution to valid assessment can be vitiated by a poorly conceptualised or constructed mark scheme. In this section we identify problems where mark schemes that do not give the marker enough, or good enough, guidance to ensure reliability and validity.

8.2.1 No mark scheme – no guidance

'Circular' mark scheme: "any suitable X"

It is surprising how many question mark schemes effectively give no help to markers:

Example:

(c) The other end B must be attached to enable it to open and close.
Explain with sketches and notes how this may be achieved. [4]

MS: Suitable drawing to enable acrylic and (b) to open and close. [4]

The mark scheme merely restates the question without any elaboration of how markers should allocate marks, what content would be "suitable", or what to do if there are no "sketches" or no "notes". If the question asks for both sketches and notes can a candidate ignore this and still get full marks? Is the "drawing" essential, as the mark scheme says?

Example:

(d) Explain why many businesses are organised in this way. [3]

MS Any valid suggestions given. [1 each]
Appropriate explanation. [1-2 each]
Maximum of 2 for suggestions
Suggestions will cover: control; communication; delegation; promotion; status; hierarchy; specialisation; etc.

This time there is at least a hint of what might be appropriate content, but it is not a complete list, and there is still no help in judging what is a good enough explanation to merit a mark.

Example:

4. The Aga Group plc manufactures a range of products used in domestic kitchens and catering businesses. Its most famous product is the Aga cast iron oven. The Balance Sheet for the business for 31 December 2002 and 2003 is shown below. Study it and answer the questions which follow.

(a) Give one example of a fixed asset which Aga Group plc might own. [1]

MS Any relevant answer BUT must be an asset not a cost (1)

Again there is no guidance on what should be accepted, though this time there *is* some advice on what *not* to accept.

We were surprised how often the phrase 'Any relevant answer' (or its equivalent) is used in these subjects. There will always be a temptation to leave the judgement of acceptability to individual markers, treating them as professionals, and we think the temptation will be greater in subjects like Business Studies and Design and Technology than in some others because they make so much use of contexts. When questions are set in the real world of business or manufacturer (and to a lesser but still significant extent in the geographical world) it will often be difficult to find a way of distinguishing the class of acceptable answers from the rest. Nevertheless, leaving decisions to the judgement of individual markers, some of whom will be inexperienced, cannot guarantee reliable marking. This issue will be discussed further later.

8.2.2 Use of ‘model answers’

It used to be commonplace for principal examiners to write, for each question, a *model answer* ‘to show what they were looking for’. This is, however, inadequate as a guide for examiners. Model answers tend to be better than can reasonably be expected from most GCSE candidates, and are rarely aimed at defining carefully the *minimum* quality needed to get any particular mark: in short, they do not give a marker any help in judging real students’ answers.

Example:

(e) Ecotourism can be a sustainable method of ecosystem management. For a named area you have studied, state fully **two** advantages of ecotourism.

MS Example of Level 3 answer:
Two stated advantages with consequences and full elaboration that in total includes two specific facts related to the chosen ecosystem e.g. Ecotourism in Kenya has helped local communities because the Maasai people have had some inward investment from tour companies, new roads, schools and fresh water supplies have been built, which increases their quality of life. Without the ecotourism revenue, such facilities would not have been built in this remote area of south western Kenya. The protected status offered for large animals in ecotourism areas has meant that some, like the lion, cheetah and even elephant have been saved from extinction. Rather than converting the land to permanent pasture – an unsuitable habitat for these animals – it is left natural, so their habitat is preserved, and the visitors only take photographs rather than hunt.

The mark scheme awards 6-7 marks for a Level 3 answer, but how good is this example? It seems far better than a borderline Level 3, so it may be of little help to a marker reading real answers. Note that this could turn into the kind of ‘best fit’ marking scheme used in language testing – including many English exams – but only if either every level is exemplified and the samples for the top and bottom levels are typical of responses in the middle of each level, or examples are given of typical borderline cases.

Example:

(ii) The amount of salmon bought from Healthy Salmon plc by German supermarkets may be influenced by the exchange rate. Fig. 2 shows the trend of the euro against the pound over the period 2003 to 2006.

Fig. 2

Using the data shown in Fig. 2, evaluate how the change in the value of the euro against the pound may have affected Healthy Salmon plc.

MS Many possible answers
Data shows
- Euro has got stronger. Fewer euros needed to by £s.

Example of a Level 2 answer
Euro is stronger against £. Fewer euros needed to buy £s. The increase in the Euro is quite significant, therefore, it is cheaper for German supermarkets to buy salmon from Healthy Salmon plc. This is likely to result in higher demand and profits. Although the price per unit of salmon not high, if they import large quantities of salmon the German supermarkets will be better off thus likely to buy from Healthy Salmon plc. However, if Healthy Salmon plc imports raw materials, it will have to spend more money as the £ has fallen which may decrease profits. [5]

Level 2 (4-5 marks)	Evaluation of data
Level 1 (1-3 marks)	Application and analysis of data

The table of ‘levels’ and marks at the end of this mark scheme is useful, but there is little help in the rest of it. The “Example of a Level 2 answer” is clearly a *very good* Level 2 answer, and markers are left to decide how close to this a response needs to be to score 5, or 4, or how much ‘analysis’ is needed to score 3, or 2.

8.2.3 Use of 'model points'

The problem of giving only model answers as a mark scheme appears in simple semi-constrained questions too.

Example:

(c) The table below shows four stages in cutting out the slot. Complete the table.

Stage	Processes
1	Mark out the slot
2	
3	
4	Glasspaper the edges of the slot

[2]

MS: Two missing stages to include: (one mark for each)
Drill hole in shape, insert blade or coping / Hegner / Scroll saw.
Routing, Mortising, Chiseling.
Saw along outline shape. File edges flat and level. **[2]**

Does a candidate need to write as much as this in each space? For example, does the last line mean that *both* 'saw' and 'file' are needed? Again, the problem is how close do they need to be to get the mark.

8.2.4 Incomplete 'Good 1' Outcome Space

Very Constrained

When questions are Very Constraining candidates must answer in the required form; here it is the name of a hardwood.

Example:

(i) Name a hardwood commonly used in the manufacture of children's toys. [1]

MS Hardwood: beech. **[1]**

But is beech the *only* hardwood that fits the question? The British Toy and Hobby Association, in a web document "produced for GCSE students who have chosen to design and make a toy", recommend beech, elm and cherry.

8.2.5 Lack of guidance for relating marks to quality

Marks for "develop" or "elaborate" or "explain" or "expand"

It is very common in some papers for marks to be divided between 'objective' ones, which are given for stating, describing or suggesting things, and 'subjective' ones to be awarded if the answer is extended in some creditable way. This extension strategy is indicated to the markers by one of the words in the list above, but is not explicitly indicated to the pupils.

Example:

(b) Describe three factors that someone taking on a franchise would need to consider when deciding where to locate a coffee shop. [6]

MS Suggestion [1 each]
Expansion/Development [1]

Suggestions might include:

- cost;
- passing trade/reference to the market/demand; (plus five more suggestions)
- etc.

Here a list of Suggestions is given (partial, ending with "etc."), but no guidance is given to help markers decide when the suggestion is expanded *enough* to merit the second mark.

A more extreme example of the same problem occurs in the next question:

Example:

(c) Astradimes believes that its coffee will sell at an average price of £3 per cup ranging from £2 in some parts of the country to £4 in London. Why does the price of a cup of coffee vary in different parts of the UK? [6]

MS Suggests reason [1]
Expands/explains/develops [1-5]

Suggestions can involve anything to do with price determination, e.g.

- cost plus;
- price discrimination;
- market price;
- demand and supply;
- skimming/creaming;
- city centre compared with outlying regions.

There is a serious issue in this question: although a list of six possible reasons are given in the mark scheme the mark allocation implies that only *one* reason should be credited, with the rest of the marks given for expansion, explanation or development. It is hard to believe that markers would use it this way, not giving credit for *both* cost-based and market-based reasons for example. But however they interpret it the mark scheme gives no help in awarding the 'e/e/d' marks, even though they carry 5 of the 6 total marks.

No rule for how good to get 2 or 1 or 0.

Another format in which the same issue frequently arises is where 2 marks are awarded with no indication of what sort of answer should get 1. The first example below is similar to the ones we have just discussed.

Example:

(c) Explain which of the two designs would be more expensive to manufacture in quantity. [2]

MS: Argue that either can be more expensive to manufacture in quantity
Design A: more parts involving more processes, more time and more costs.
Design B: the brackets would need to be joined to a wall plate.
Award 0-2 marks dependent upon quality of explanation. [2]

First, we see here another problem with 'explain'. Is it right to ask for an *explanation* for something which is uncertain? The mark scheme accepts that *either* design "would" be more expensive, but this is logically inconsistent, as they cannot both be. If either is to be acceptable the command word phrase is at fault for implying that one of them is the right answer; it should have asked "Explain which … you think would be more expensive" or "Which do you think … ? Explain/justify your answer" or even just "Which do you think … ? Why?"

But the reason for quoting this example here is that it instructs markers to judge the quality of the response in awarding 2, 1 or 0 marks but gives them no help at all in deciding how much 'quality' (that is, how much of the trait) needs to be evident for each level of score.

0/2 or 0/3 scoring

There is normally an understanding in points-based examining that each mark will be awarded for some specific point, but in some cases it is hard to see how there are enough points in the issue to justify the number of marks available.

Example:

(iii) Explain two reasons why a high level of working capital is more important in a business than a high level of start-up capital.

1. .... [2]
2. .... [2]

MS Working Capital is more important:

- for the day to day running
- to pay bills, wages and expenses

(2 × [2]) [4]

The two points in the mark scheme in the next example are quite simple to state, and there is no indication that anything more than a mention of each is required – despite the command word asking for an explanation. How then would an examiner decide to award 1 mark rather than 0 or 2? The question appears quite complex, and might justify four marks if answered and assessed properly, but it is not clear that this mark scheme can do it justice.

Where the response format is not, or not entirely, verbal it is perhaps more understandable that the mark scheme doesn't clearly specify criteria for partial credit, but it is not at all clear what markers should do with an instruction like the next one:

Example:

(a) Complete the sketch below to show how the 6mm thick panel could be fitted to the frame of the easel at A.

MS: Completed sketch to show appropriate method: rebate, groove or beading. (3)

In the absence of positive criteria for 1, 2 and 3, there is a risk that markers may adopt a negative approach, penalising each fault that they find with the completed sketch.

8.2.6 Multidimensionality of 'levels' mark schemes

We have looked at examples in which examiners have used 'points' to assess the quality of answers. But there are also many examples where rating scales are used to address quality directly. The next examples show a problem that often occurs.

Example:

iv) Cliff recession causes many problems for people who live in coastal areas.
Choose a case study of a stretch of coastline or coastal area that is suffering from cliff recession.
Chosen stretch of coastline or coastal area.
Explain the causes and effects of cliff recession in this area. (5)

MS Levels mark

- Walton case study
- Section 1A of appendix

Do not credit management

Level 2 5-4	Specific detail of a case study must be included to reach level 2 There should also be explanation of causes or effects
Level 1 3-1	Descriptive comments only about causes and/or effects. Likely to be very general. Not related to case study.

(5)

The mark scheme gives some rather terse detail about relevant content, and the advice to *not* credit 'management'. But marking is to follow the levels scheme provided. The problem is that there are at least two clearly different qualities included in these level descriptors – use of a case study, and explanation/description of causes or effects. How should a marker reward an answer that is very good on describing a case study, but fails to explain well? It is clearly Level 2 on one quality but just as clearly Level 1 on the other. It is likely that different markers will choose different strategies for such a case; effectively they will be assigning different weights to the two (or more) dimensions of quality.

This example comes from a Foundation paper. Curiously, the same question was also set in the Higher paper; the only difference was in the mark scheme:

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.

(5)

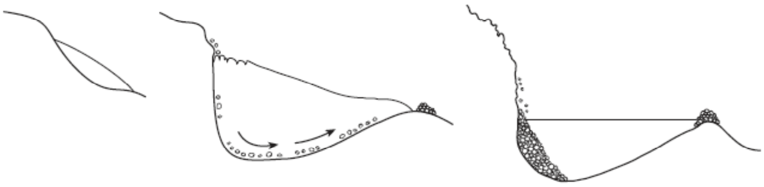
The top descriptor for Level 3 in Higher does not seem clearly different from the descriptor for Level 2 in Foundation. There does not seem to be any reason why different schemes should be used in the two papers.

In the next example the examiners have recognised the need to describe two dimensions separately.

Example:

(iii) Explain the formation of a corrie and its lake.

Add labels to the diagrams below to illustrate your answer. [6]



Space was provided for writing as well as for labeling the diagram.

MS

Level	Explanation	Communication
Level 1 1-2 marks	Some understanding of the nature of corries. No or irrelevant labelling.	Information is communicated by brief statements.
Level 2 3-4 marks	Understanding of role of glaciers and can name processes. Basic labelling. NO LABELLING = LEVEL 2	Communication may be verbose or illogical. A limited number of specialist terms are used. There is some accuracy in spelling, punctuation and grammar.
Level 3 5-6 marks	Clear explanation of abrasion and plucking / role of deposition / tarn formation. Clear and well labelled.	The written style has a suitable structure. There is a range of specialist terms. Spelling, punctuation and grammar have considerable accuracy.

The distinction between Explanation and Communication is very clear, though the Explanation scale may still be a mixture of the quality of the labelling and the quality of the explaining, but there is still a real difficulty if a pupil ‘explains’ at Level 3 and ‘communicates’ at Level 1. Should the marker compromise and award a Level 2 mark, when the pupil’s response is in no way typical of Level 2? Or should they decide which of the two scales is the more important as evidence of the pupil’s mastery of the trait of geography achievement?

This kind of ‘best fit’ rating scale system is appropriate for National Curriculum assessment if a pupil has to be assigned to a single level, but the risk of invalidity that it allows through unreliable marking means that it is not appropriate for rating the answers to examination questions.

8.2.7 Inappropriate use of ‘points’

Points-based mark schemes are typically used with Semi-Constrained and Un-Constrained questions, where pupils are given the opportunity to express their own idea of the answer, or are faced with the demand of constructing their own response. But there can be difficulties with writing the mark scheme, or rather with detailing the outcome space, even for questions that are close to the Very Constrained end of the spectrum. Consider the idea of a question – ‘What does *deforestation* mean?’. We found this question in three papers from two boards. One board asked, in both tiers:

Example:

(i) What is the meaning of the term ‘deforestation’? (1)

The mark schemes were not the same in both tiers:

MS:

Removal of trees (HT)

Removal of trees, cut down, burning, destroyed. (FT)

The differences are interesting. It seems that at Foundation a pupil can get the mark by mentioning any reference to destroying a tree in any way, although the mark scheme is not clear about whether destroying *one* tree alone gets a mark. At Higher it seems that more than one tree must be removed, though no rule is given about using specific terms like ‘burning’. It may be quite appropriate to mark the question differently for different groups of candidates in this way, but in both tiers markers would have been helped by a clearer principle to distinguish between what is and what is not good enough to get the mark.

The other board developed the idea differently:

Example:

(i) State the meaning of the term deforestation. (2)

MS Level 1 ([1]) An incomplete definition

- Cutting down a tree
- Removing trees

Level 2 ([2]) A full definition needs some reference to scale

- The complete clearance of a forest area by cutting down or burning trees

The only difference in the question paper, compared to the previous board, was that 2 marks were given rather than 1, but the mark scheme is very different. Even for just a 2 mark question it is a ‘levels’ scheme, with definitions and examples for each level, rewarding a variety of quality that the other examples could not. The example shows how much scope there can be to develop a simple idea into a more efficient measuring tool, even though there was nothing really wrong with the first board’s questions.

But points-based schemes may sometimes be seriously flawed.

Example:

(d) New materials, such as carbon fibre, are used extensively in the construction of bicycle frames.
Explain **two** advantages for the user of a bicycle of having a frame made from carbon fibre.
1 ....
2 .... (4)

MS:	4	(d)	**Two** advantages explained:
a	*Single piece frames* therefore *no weak spots/greater rigidity*	2x1	(4)
b	*Lighter frames* therefore *can go faster/easier to carry*	2x1	
c	*Good strength to weight ratio* therefore *frames can be slimmer/slighter*		
d	*No surface treatment/finish* since *carbon fibre does not rust*		
e	*Stronger material* which can therefore be *subjected to greater forces when mountain biking*		
f	*Can be moulded* which means *complex shapes can be achieved*		

(The letters on the left of the mark scheme have been added.) In this example, each italicised point can earn two marks if suitably explained. A pupil giving answers *b* and *e* can therefore score 4 marks. A pupil giving answer *c*, however, can only scores 2 even though this answer is clearly more sophisticated than *b* or *e* – in fact it is equivalent to both of them together.

A ‘points’ mark scheme rewards quantity rather than quality and is incapable of giving more credit to better answers with the same number of points.

Example:

(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 many Semi-Constrained questions like this, where an *explanation* is asked for, and the answer is a sort of sketch of a model answer with creditable points indicated. The problem is that the ‘model’ contains 9 points for the 4 marks; it is explicitly stated that only four can be given credit. But any four? According to the mark scheme the answer below should get full marks:

> “Gravity makes the glacier move slowly, and over time it smoothes the surface.”

Is it adequate? It makes no mention of the *essential mechanism* of abrasion, which is the scraping of the ground by rocks dragged down by the moving ice. It seems quite invalid to give 4 marks to someone who offers an answer that has no mechanism to explain a physical process. Yet each of the nine points is worth crediting as a fourth mark if the three essential points are also included. The problem 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.

The attraction of a points scheme seems to be that, by turning marking into a fairly objective process of spotting the points students mention, it maximises the reliability of marking, but the price may often be the loss of validity.

Rather than seeing this simply as a fault in the mark scheme we could consider it as an attempt to use the wrong kind of mark scheme for the question, and we turn to this issue of mismatch in the next section.

8.3 Mismatch between the question and the mark scheme

It is not enough to write good exam questions that ensure the students' minds are doing the things we want them to show us they can do; our validity principle demands that we also give credit to, and only to, the evidence that they can do these things. We saw many cases where we think the mark scheme would not help markers achieve this, even though it might seem on first reading to be good, and where there was therefore a risk of validity being compromised. In this section we concentrate on cases of mismatch between the task set by the question, what the students' minds are likely to have been doing in response, and how credit was awarded (at least according to the mark scheme).

8.3.1 VC, C, UC with inappropriate mark scheme

We begin with an important principle for the design of mark schemes: the location of a question on the scale from Very Constrained to UnConstrained largely determines the type of mark scheme that is appropriate.

If students are very constrained by the question, then **Good 1**, the set of acceptable answers, may be easily described as a single number, word or phrase, or by a short list. Even a slight loosening of the constraints, however, may mean that this is inappropriate.

'Points' or 'Right/wrong' where Generic seems more appropriate

Consider this example:

Example:

(ii) Advise Cafedotcom on **one** other way it could segment its market. [4]

MS Any other market segment with advice, e.g.
pensioners
school children (*etc 9 segments listed*)
Do **not** accept gender only
(1 × [4], i.e. [1] for identification, [3] for example [4]

There is no guidance at all on how to award the 0-3 marks for "advice"; indeed, the mark scheme gives [3] for "example". It would seem much more appropriate here to use a 'levels' type of mark scheme to judge directly the quality of the advice offered.

Example:

(c) Should Peter stay as a sole trader, rather than taking on a partner, when starting up
Pete's Pantry? Discuss **both** options when giving your recommendation. [8]

MS One mark for each relevant point. Maximum of 6 marks if only one option is addressed.
One mark for each advantage and disadvantage of the options stated and explained. Reward also conditional statements and development points. Do not reward "mirror" arguments.

- Sole trader keeps all profit (1) so will have more money for himself (1), retains control (1) so does not have arguments with partner (1), is easy to set up (1) because he does not need to draw up a Deed of Partnership (1).

(*etc. four more paragraphs like this*)

- I would advise him to set up as a limited partnership (1) so that he gets the advantage of limited liability (1). Max: 8 marks

After the first two paragraphs of general advice, the mark scheme turns into a model answer with points indicated. (See earlier for a discussion of model answers.) There is an implication that the advice *must* be to take on a partner, since that is the only decision that appears in the mark scheme. But it seems wholly inappropriate to score this question by counting relevant points. As in the previous example, would it not be more sensible to use a 'levels' type of mark scheme to judge directly the quality of the recommendation made?

A third example shows how a poor mark scheme may fail to address the question actually asked.

Example:

(a) Peter and Rosie plan to sell sandwiches, cakes and drinks from Pete's Pantry. They designed the following advertisement.

> **Pete's Pantry – Opening June 30th**
>
> Sells good grub. { graphic }
>
> Come and try us – first cake free.

How successful do you think this advertisement would be? Give reasons for your answer. [4]

MS (a) Target: Ability to evaluate the appropriateness of an advertisement.

One mark for each point of criticism and one mark for an explanation of why it is a problem.

- No address/give the address (1) – customers will not know where the shop is (1).
- No contact phone number/give the phone number (1) – customers will not be able to contact the shop to make orders (1).
- More details about the food sold (1) – so that customers know what is on sale and will be more likely to be tempted to buy things such as sandwiches/cakes (1).
- More details about prices (1) – so customers can compare prices with other business (1).
- What time the shop opens (1) – so customers know when they can go to the shop to buy things (1).
- More pictures (1) so that it looks more attractive to customers (1).
- It is good because it offers a free gift (1) and customers like something for nothing (1).
- Make it more colourful/use more pictures (1).

- The word "grub" does not create the right image (1).

NB 2 marks maximum for positive comment.

List = max 2 marks. Max: 4 marks

The task phrase is "How successful do you think" which requires an evaluation of the advert. The mark scheme, however, is a kind of critique of it, with suggestions of how to make it better, but no overall judgement.

Poorly defined Generic scheme

In contrast to the previous examples, some mark schemes that do aim to assess responses in a generic way make the error of ignoring the content aspect of the outcome space.

Example:

(b) Use notes and sketches to show how a simple mechanism could be used to make the hopper tip.

MS: Practical Mechanism (0-3 dependant in detail) (0-3)

Quality/ clarity of communication (0-2)

When the examiners wrote and reviewed this question they must have imagined the sorts of mechanisms they would see in pupils' responses. They must have had a few ideas of mechanisms they would consider as 'good', and probably of one or two they would expect to see that would be 'poor' and not work. That is, they will have had a more or less clear idea of the expected outcome space. It would have helped the markers considerably to have had a list of 'good' and 'poor' mechanisms as a 'crib' sheet when marking: instead, they are required to study each answer individually to see if it is 'good' or not. In practice, each marker will create their own crib sheet, in memory or on paper, as they mark more and more responses: *to the extent that these differ between markers the process will lose reliability*.

The criterion for 'good' is stated to be "Practical". It is not obvious to us if this simply means workable, or whether it includes economic and aesthetic considerations.

An ideal mark scheme for this kind of question should map out the expected outcome space by listing the most likely 'good' and 'poor' mechanisms, and define with a bit more care the principle of practicality that should be applied in unexpected or doubtful cases. After that the assessment of quality as "dependant in detail" can be applied.

8.3.2 Misleading command word in question or mark scheme

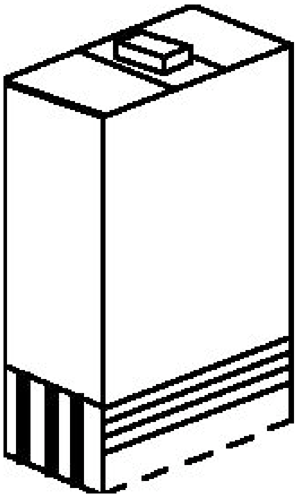
The command word is a direct instruction to the pupil, telling them what kind of response to make. It is surely important that the mark scheme gives credit to pupils who do what the command word says they should do, and only gives full marks to those who do it well.

In the next example, the command word is “annotate”:

Example:

(iii) An earthquake will usually destroy low buildings but not taller buildings.
Annotate Figure 2d to explain why.
You may add to the diagram.

Figure 2d



(There was plenty of space around the diagram for arrows, labels or other annotations.)

MS Point mark
Rubber foundations to absorb tremors
Counterbalance on roof to restrict swaying
Steel framework to allow for swaying (3)

What is the actual instruction here? To ‘annotate’ normally means to ‘add explanatory notes or critical commentary’ to something, so the idea of annotating the diagram to explain something should be clear enough. But, if so, then why do pupils need to be told that they “may add to the diagram”? Annotating without adding is not possible. It seems that the examiners were not confident that pupils would understand what they were supposed to do – what ‘annotate’ really meant – but the extra sentence, rather than allowing them to add notes or labels, might suggest to some pupils that they should draw extra details on the building in the diagram.

In general, it is not wise to use command words that need extra glossing like this.

“Explain” marked as something else

‘Explain’ is one of the most common command words in GCSE exams. But it is often misused, and the marks are given for descriptions, definitions or even just for naming something.

Example:

(i) Explain what annealing does to the mild steel rod. [1]

MS: Annealing softens the metal, makes it malleable, easier to bend [1]

In this case the ‘explanation’ is nothing more than a description of the effects of annealing a rod.

Something else marked as "Explain"

The reverse also happens:

Example:

(f) At present Wholesome Drinks Ltd uses the following channel of distribution for its goods:

Manufacturer → Wholesaler → Retailer

Identify the advantages to Wholesome Drinks of using this distribution channel. [4]

MS Two advantages from:

- Wholesaler buys in … (*list of four 'advantages'*)

([2] for each advantage explained) [4]

In this case there is no indication to the candidates that any explanation is required, and the command word phrase "Identify the advantages" – giving no indication of the number of 'advantages' to be identified – strongly implies that none is.

8.3.3 Implicit/hidden structure in the mark scheme

Students will use (though some will miss) any hint they can find to help them structure their answer. An important source of clues is the number of marks, taken together with any quantitative words in the question. If there is a mismatch between these, or between them and the mark scheme, unfairness is a likely consequence.

Between points

Example:

(i) Describe how CAD/CAM could be used to help in the design and manufacture of this notelet holder. [2]

MS (i) Must include ONE CAD and ONE CAM reason -
CAD used to draw out the shape of the net accurately or a 3D image I .
CAM used to cut out the outline shape and engrave the bend lines, CAM will ensure repetitive control I [2]

It is not obvious that the oblique in 'CAD/CAM' stands for 'and' rather than 'or'. Nor is it obvious that the answer must be at a low level of detail instead of describing in general how IT can help reduce costs or speed up manufacture. It would seem quite reasonable to give one well developed answer rather than two brief ones, one from each of CAD and CAM. Pupils cannot assume that 'Describe how' *necessarily* requires a description of a process, since it is often used in other ways, like as a synonym for 'explain'.

Example:

(ii) What problems might arise in Cafedotcom if the partners do not draw up a "Deed of Partnership?

____(*8 lines*) [4]

MS Two problems with brief outline, e.g.:

- distribution of profit and loss etc (*a list of 7 such possible 'problems' is given*)

(2 × [2])

There is no hint in this question that the mark schemes requires candidates to identify exactly two specific problems, rather than one, or three or four, or to give a more abstract answer in terms, perhaps, of distribution of finances or breakdown of trust.

Between Assessment Outcomes

In some exams, notably in Business Studies, it is common for a question to be scored on two scales which are derived from the Assessment Outcomes given in the syllabus test specification, but the way it is split is generally not made clear to the candidates. A 4 mark question may be split 2 + 2 between two assessment outcomes, a 6 marker may be split 2 + 4 or 4 + 2 or 3 + 3. The example below shows an 8 marker.

Example:

(b) Discuss whether penetration pricing would be the most appropriate pricing strategy for Center Parcs to use for its new conference business customers. *(8 marks)*

MS Possible areas for discussion:
(*list of five areas*)

	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)

The command word "discuss" may be understood to involve evaluation, but it is not made clear to the candidates that half of the marks are to be for the judgement they make. In this case, in addition, it is hard to see how the quality of the judgement could be graded into four (actually five) levels of quality.

8.3.4 Confusion between specific and general

In a few questions it seems that a generic points-based mark scheme was being used, perhaps re-used, without modifying it to suit the specific question:

Example:

DEDUCTIONS	
PAYE	£63.00
N.I.	£44.00
PENSION	£26.00
CHARITY	£2.00

(e) Give **one** example of a voluntary deduction that Jane pays from her wages. [1]

MS Any one from the following:

Charitable donations	Subscriptions to trade unions
Pension contributions	Insurance
Named relevant examples	etc.

(1)

Observant candidates will have noticed that there are only four deductions from Jane's wages, *not* including a trade union subscription; are others to get the mark for writing "Subscriptions to trade unions" even though it is untrue? And how many trade unions might Jane belong to? Clearly the mark scheme addresses the general question of voluntary deductions any waged employee *might* pay, and not the question asked.

8.4 Other issues

There are a few very general issues within one or other of the subjects that we felt are important enough to be highlighted as issues for consideration; although they arose in one subject they are not necessarily limited to it. We can only describe the problems caused by present practice; we cannot rule on what *should* be done since these all arise from a conflict between two or more worthwhile principles.

8.4.1 Case studies: real or invented?

Geography examiners are well aware of the seemingly inevitable problem they face with Ordnance Survey maps – that any bit of the United Kingdom whose map they use in the exam will be more familiar to some pupils than to others. Although they could, in principle and at considerable expenditure of ingenuity, time and money, draw their own map using the OS principles of design they prefer to use real maps for reasons of authenticity as well as convenience.

A similar choice faces examiners in Business Studies. All of the questions in all of the papers we looked at are contextualised (although the context is sometimes not used in a specific question). In one board **only** real businesses are used: we saw Center Parcs, Marks & Spencer and the Ford motor company. In another board four out of five cases each year were real, including Aga cookers, Daimler’s Smart car, Valleywood Studios, and Rachel’s Organic; the others were probably invented. The other boards only use simulated cases.

Does this matter? We saw two reasons for concern. First, as with Geography, there is the issue of unfair knowledge. The Ford example mentioned that the company employs nearly 30,000 people in the UK; it makes a considerable impact on the life of many more in the areas around its main sites. While every pupil will know something about Ford, some will know much more. A potentially more serious version of this issue comes from the habit in one board of picking their text for real cases from the BBC website, since this means that the issue highlighted must have been significant enough to attract attention from the national press. The case of Howies, a manufacturing company with five staff in West Wales, being sued by Levi Strauss for copyright infringement in 2003 certainly did ‘hit the headlines’, and is likely to have been used extensively in teaching Business Studies in the following few years. Was it wise to use it in an exam in 2006? Was it fair, if some classes had discussed it when others had not?

Geography can suffer from the same ‘headline effect’, as this example shows. It is based on the Indian Ocean earthquake and tsunami of 26th December 2004:

Example:

(iii) What caused the loss of life in the affected countries? (1 mark)

Appearing just 17 months later, it is likely that many pupils will have answered this question by remembering the news coverage rather than by trying to apply geographical principles. It was also a highly emotional event and recalling it in an examination might distract some pupils from the mental activities about which the examiners want them to show evidence.

The second concern relates to the fit between the available information on a real case and the aims of a question writer. The purpose of a question is to elicit evidence that a pupil knows or does not know things, or can or cannot carry out activities that are specified in the syllabus. An authentic case may not address the issues in the way that examiners would like, or the information about how the real company actually operates may not be publicly available. How, for example, can an examiner base questions on employment strategy on a real case if the company does not publish its strategy? There must be a temptation, at times, to invent details, or at best to ask about how companies “like” the real one *might* operate, which rather defeats the point of using an authentic case in the first place.

The case of the Smart car has already appeared in this report. The text for it was “adapted from www.bbc.co.uk” and the adaptations are revealing:

Original	Adapted
The Swiss entrepreneur behind the Smart car, whose product was delayed after it failed the notorious "elk test," have admitted that he and his partner Daimler-Benz, may have rushed its development in their haste to launch it by next April. Speaking on German television, Nicholas Hayek of the Swiss company SMH, admitted that the delay until next autumn would cost the joint venture some 300 million deutchmarks (*sic*). Caroline Wyatt reports from Bonn.	In the late 1990s Nicholas Hayek, a Swiss businessman, worked with the well-known luxury car manufacturer Daimler to finance and produce a vehicle which was sold under the Smart car brand.
The two-seater Smart car is the latest attempt by Daimler-Benz to muscle its way into the small car market, the industry's biggest-selling sector.	The low price, two-seater Smart car was Daimler’s attempt to enter the small car market and to break out of the luxury market.

The removal of irrelevant information from the first paragraph unexceptional, but the addition of the phrase “luxury car manufacturer” was a deliberate steer to the pupils; the company make other vehicles too, such as vans and trucks. In the second paragraph the word ‘sector’ was removed, wisely, as it certainly have interfered with the notion of ‘segment’, but there is less reason to add the phrases “low price” and “break out of the luxury market”. It seems that the examiners wanted to push the pupils towards a stance that Daimler wanted to diversify from the rich people’s market into the much larger moderate or poor people’s market, and this is presumably why they asked Question 3a:

Example:

(a) Suggest and explain the market segment at which the Smart car was aimed. [2]

MS: Suggestion (1)
Explanation (1) [2]
Suggestions must be specific might include the following:
gender – men/women
age – old and young
income – rich and poor
socio-economic group – class
lifestyle points e.g. number in family
etc.

The mark scheme is completely generic, as some of these points are not mentioned at all in the text, and it gives no guidance as to what the ‘right’ answer is. Yet, being a ‘real’ case, there must be a true market segment that the car was aimed at. The strong impression, from the adapted text, is that it was aimed at people of moderate means who could not afford a luxury car. The truth, whatever it may have been, is different now:

> “All that - plus a sticker price starting at $13,000 - has helped the company snag the youngest average buyer of any global auto manufacturer, a snappy 37. And Smart's buyers are an enviably affluent bunch. Nearly half pay in full and in cash.” (*Wired*, Oct 2004)

The examiners seem to have distorted the truth of their case, in order to simplify it to suit the questions they wanted to ask in 2006. But is it right to promote false facts like this? And how many pupils, car fans or merely trend conscious teenagers, would be confused by the false message?

Authentic contexts, used accurately, can bring many benefits to teaching and assessment, but it might be better to invent cases that are similar to real ones if the truth doesn’t fit the examiners’ needs.

8.4.2 Pre-release

A second worry concerning Business Studies concerns the pre-release policy. Four of the boards release information relating to a case study that will be used in the exam, but one does not. Of the four that do, the nature varies:

	Status of pre-release	Other written papers in the examination	Nature of pre-release case	Length of pre-release document	Date of pre-release (June 2006)
AQA	Compulsory paper	No other paper	Real case	9 pages without detail	10 March
CCEA	Compulsory paper	1 other paper	Invented case	12 pages with detail	1 March
Edexcel	Compulsory paper	No other paper	Invented case	1 page without detail	3 October
OCR	Optional paper	2 or 3 other papers	Invented case	10 pages with detail	1st January
WJEC	No pre-release	1 other paper	-	-	-

There are several differences between exams, as the table shows, and the wisdom of each might be debated. The issue that concerns us most for its relevance to the quality of exam questions is the influence of lengthy exposure of the content material, with or without details, in the public domain before the exam takes place.

We believe that, in an ideal examination, all pupils will have had equal preparation from their teachers or elsewhere, and that their minds will then be benevolently manipulated by the questions to elicit the best evidence of their achievement. Of course teachers are not all equally good at preparing pupils, but they are all professional and in contact with the examining board. But there are now alternative sources of information for pupils, including internet sites purporting to give you "the best preparation for an exam paper based on the pre-released material". For example, here is an endorsement of one such web-site from a grateful pupil:

> *Attached Files*
> *Case Study – 'EPP OBJECTIVES' –.doc (80.5 KB, 71 views)*
> *Excel Pictures plc4.doc (73.5 KB, 50 views)*
> *PreetCasestudyEPPRevision.doc (126.5 KB, 48 views)*
> *What is in the syllabus EPP Rev.doc (95.5 KB, 46 views)*
> *Financial Efficiency-Ratio Analysis.doc (26.0 KB, 43 views)*
> *meep...exams :S*
>
> ***Re: Edexcel Case Study Excel Pictures plc (EPP)***
>
> *Thank you soooo much My teacher is rubbish and I don't know n e thing about my case study so this was perfect...if u ever need something on n e other subject (other than business studies lol) just ask*
>
> *Thanks again, Maria*

Another web-site offers teaching courses to prepare pupils for the AQA and Edexcel pre-release papers, indulging in a quite remarkable exercise in question spotting, given how little detail those boards include in the released document. The issue is not whether web-sites are better or not than teachers, but simply that the appearance of such uncontrollable 'resources' makes it ever more likely that examiners will not be able to influence the pupils' minds in the ways that validity requires.

8.4.3 Pictures and other resources

In a Geography exam pupils have to handle, more or less simultaneously, a remarkable number of separate pieces of paper, diagrams, photographs and maps. When you add the routine ways in which emphasis is used, this can lead to questions that look rather complicated:

Example:

(a) **Box A** on the **OS map** and **Figure 1** on **page 2** of the **Resource Folder** both show a river entering the sea.

We presume that pupils are well drilled in using multiple simultaneous resources by their geography teachers, and that the ability to handle this kind of complexity has been judged to be a reasonable demand, a valid part of the trait being measured. Nevertheless, we are concerned when resources seem to be used merely for form or out of habit.

There were several occasions when we felt that photographs, in particular, were included merely for decoration or because pupils would expect to see them. For example:

Example:

(a) Look at Figure 3a. Also look at Photograph D in the Map and Photograph Booklet. These show tourist activities on a beach in Cuba.

(i) **Active** or **passive** is one way of classifying tourists.
Use the words **active** or **passive** to complete the labels on Figure 3a.

Photograph D shows a beach scene in which some tourists are sailing, using pedalos, and sunbathing. Figure 3 is a sketch of the photograph, with labels pointing to one example of each activity to be completed with ‘active’ or ‘passive’. The photograph adds nothing at all to the question, except by increasing the demand for handling complexity; it is simply an extra distracting resource and another bit of paper for them to locate and process. It is true, in examining as elsewhere, that photographs are far more powerful than words; they activate far more ideas than the examiners want activated, they distract and they may easily mislead. They should be used with care.

The issue of handling multiple resources simultaneously is likely to become more salient as the use of IT in examining spreads. One of the most valuable contributions IT may make is to facilitate the use of multiple and very large data resources, and the process of integrating them into a single response document. All three of our subjects may benefit from this.

9 Conclusions

In Section 8 we concentrated on the problems we saw, since our remit was to look for ways to improve the quality of GCSE assessment. As you have seen, we classified the problems into three groups:

(i) writing questions that will ensure the students' minds are doing the things we want them to show us they can do;

(ii) writing mark schemes that will help markers;

(iii) ensuring that the mark scheme matches the question, so that we give credit when the students do the things we want them to show us they can do.

In these conclusions we make suggestions for ways to avoid, overcome, or at least manage, the problems we identified. We are recommending a more integrated and theory-based approach to question and mark scheme writing. Section 9.1 deals with writing the question and mark scheme together. Section 9.2 looks in more detail at the principles underlying different types of mark schemes. Finally, Section 9.3 deals with validity – so that every question contributes to the overall measurement of the trait that we want to assess.

9.1 Using OSCA theory to create exam questions

We believe that the single most useful contribution we can make to the improvement of GCSE examining is to propose a system for creating questions and mark schemes that will make it easier to elicit and evaluate the evidence examiners want to see. OSCA is our way of doing this.

As described earlier, Outcome Space Control is the core of Assessment practice – the ability to discriminate depends crucially on the range of responses that students make to a question. The task of a question is to elicit an appropriate outcome space, and the task of a mark scheme is to evaluate that outcome space in a way that accurately represents the trait the exam as a whole is intended to measure.

9.1.1 Why OSCA is different

In traditional assessment theory, a theoretical, abstract domain of knowledge is assumed, from which an examiner can sample pockets of knowledge to assess. The examiner then creates an item to assess that knowledge. Assessments are designed to tap into a latent trait in students – called ability in the subject. This ability depends largely on knowledge and skill in the subject, and the magnitude of a test score is taken to represent the amount of the trait in the student. Question writers are interested in discriminating between those who know the correct or best answers and those who do not. Question clarity is essential to ensure that the scores relate to the latent trait, and problems arise when examiners would like to assess a particular piece of knowledge that is not easily translated into an item. The assessment itself can be evaluated by pre-testing items and investigating the statistics to find out whether there were general misinterpretations, how difficult the item was and so on. Assessments might be designed to discriminate between students and therefore need to combine to produce a good spread of scores on the test overall. Traditional assessment theory is largely concerned with these operational properties of tests.

In contrast, OSCA theory focuses upon the possible set of student behaviours. These are observable, as opposed to the abstract, theoretical domain of knowledge in traditional approaches. Since student performances are what is being considered, there is no place to discuss knowledge that an examiner would like to assess that is difficult to demonstrate. If it cannot be demonstrated, it is not part of the set of possible assessable elements.

With OSCA we do not set out to test the domain of knowledge, and exercise our ingenuity to minimise all other skills and demands. Rather we conceptualise the trait we want to measure as *an appropriate blend* of all the knowledge, ability, skills and demands that we think constitute the kinds of behaviour we want students to show us. We aim to design questions and mark schemes that will elicit and measure both good and poor levels of performance, making sure that the skills or abilities that differentiate these performances are appropriate for our purpose. The exam is designed to elicit behaviour which is visible and directly available to the examiners – rather than to measure ability which is always latent and can only be inferred from the visible evidence. To report a student's ability in a certificate is to make another inferential leap from the evidence to what it implies. This inference is, of course, very important: exams will not be useful if we cannot generalise from the score on one exam to how the student is likely to behave in other contexts. That is why we stress the importance of ensuring that the behaviours the students are engaged in during the exam, and for which they are rewarded, do properly reflect the overall trait we want to assess.

Question writers attempt to guide the student to produce particular kinds of answers. This means more than writing questions clearly, it means signalling what kinds of performances will be worthy of credit. Evaluation of the assessment itself in this approach tends to be based upon craft knowledge of the kinds of effects that the assessment might have upon students' performances. Statistical information regarding the difficulty of items and so on is less crucial in this approach because OSCA concentrates on the demands that make up the test trait, rather than its difficulty. Demand and difficulty can be distinguished conceptually; the demands of a task are considered to be the same for everyone, even though some students may find it easier to meet them than others do. By deliberately designing various demands, in various amounts, into the tasks examiners determine the nature of the trait they want the test to assess.

OSCA is concerned less with the general operation of the test than with the interaction between the test and student taking the test. In this approach, the test and its items are communication devices between the question setter, the students, and the marker.

9.1.2 The OSCA model

In recent years our programme of research in these areas has led us to suggest an 'ideal' model for writing an exam question. In this we tried to facilitate the production of not just a question but also the mark scheme that would optimise the quality of the whole assessment. During this study we have refined the model further, and developed our OSCA theory, based on what we have seen by studying questions and mark schemes together with the syllabuses and their intended assessment outcomes.

An exam question 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 how well they can do it.

1 We begin with the **idea** of a task. This is essentially a creative process and very difficult to control. In 2002 Haladyna, who has probably made more effort than any other academic to systematise question writing wrote:

> 'Toward a Technology of Test-Item Writing' *(Roid & Haladyna, 1982) reported on the status of item-writing theories current to that time. None of these theories survived. … With the publication of these essays, theories, and research, the scientific basis for writing test items appears to be improving but very slowly.* (Haladyna, Downing, and Rodriguez, 2002)

It seems that the initial steps in question writing will remain essentially a creative process for the foreseeable future. In Section 9.1.4 we do, however, discuss how examining bodies might encourage creative questioning in their examiners.

2 At this point, we suggest writing down carefully not the question but the **key idea** that the question will try to address: the basis for evidence that will discriminate between those pupils who are better at the subject and those who are poorer.

We suggest leaving the question itself quite vague, and turning instead to the **outcome space**. Remember that the outcome is the only evidence that we will get on which to decide how much ability or knowledge the student has. What kind of evidence would we like to get using the idea we have? It should be possible for the "question writer" (this term is now being used as a shorthand for someone whose task is more than just writing the question) to describe explicitly or qualitatively the differences in outcome they would like to see between students with high or average amounts of ability, or between those with average and low amounts. It is this difference that will allow us to claim we have assessed the abilities of the students validly.

3 A first draft of the **marking scheme** can now be written, using an appropriate type of scheme. In Section 9.2 we describe in more detail a taxonomy which will help examiners choose the kind of mark scheme that is most appropriate.

The purpose of the mark scheme is to guide markers on how to distinguish two or three or more levels of performance amongst the intended students. It must ensure validity by awarding more marks to students whose responses show more or better evidence of the trait. This is discussed further in Section 9.3.

To be helpful to markers assessing real responses a mark scheme should not limit itself to the expected 'good' answers, as in the traditional and useless 'model answer' approach. For any question that is not perfectly objective, markers will have judgements to make to decide whether a response is good enough to get a mark, or an extra mark. To help with this the mark scheme should describe or list answers that are not good enough, and pay particular attention to defining the boundary between what is worth, say, 2 marks and what is worth 3. This is best done by a rule rather than just by listing examples.

Since the question is to be part of an exam assessing the overall trait of achievement in the subject, it is often valuable to show how you expect the trait to be realised in the range of students' responses to the question: what you expect to see as evidence for better or poorer achievement.

4 Now, and only now, we suggest drafting the **question**. The purpose of the question is to elicit from the students the evidence that the desired outcome space describes; the task for the question writer is to craft it in such a way that it does this. Students' responses will only constitute valid evidence if their minds are doing the things we want them to show us they can do, and it is essential therefore that the question makes it clear to them what they are supposed to do.

This places a high demand on the question writer's linguistic skill. Advice on language in exam questions is fairly familiar, but is not always expressed with sufficient care. For example, we have seen the rule 'Always use simple sentences' in more than one source, and we have seen dreadful questions written as a consequence of it. The rule is ambiguous, since 'simple sentence' is a technical term in linguistics meaning a sentence with just one finite verb, as opposed to a 'complex sentence' with more than one. The rule *ought* to be: 'Always use easy natural language', since the intention is to ensure that, as language testers sometimes express it, 'the question does not get in the way'.

Particular care needs to be applied to choosing the command word or words, since these carry the direct instruction to the student. We will discuss these further below.

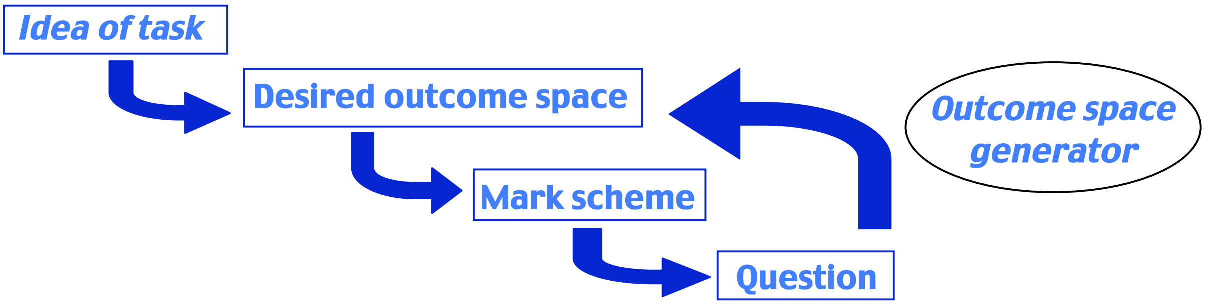
5 Once a question has been drafted we suggest that a second person follows our **outcome space generator** procedure, looking for any possible misunderstanding of the task (if any misunderstanding is at all possible it is very likely that some student somewhere will make it) and predicting the responses that a range of students will make. Of most concern will be the candidates we characterise as naïve, anxious, borderline students, since the stressful context of a high stakes examination will affect them most of all. For these students, it is never acceptable to think "They're 16, and old enough to understand what we mean".

This second examiner should discuss with the first any problems the question might pose, and the two should modify both the question and the mark scheme to ensure that the question is likely to elicit from the target students the range of responses that are catered for in the mark scheme.

It may be that there should be three people involved, the third being a language specialist and subject non-expert: only a chemist would not think it strange ask "Which fraction has the highest percentage?".

6 After the exam has been taken there may still be a revision and **finalisation** of the marking scheme – coordination and standardisation – but this ought to involve less change if questions have been written in this way.

To summarise, we propose this process:



There is a loop back from the question to the desired outcome space to ensure that the question writers begin with, and always keep in mind, that the purpose of the question and mark scheme is to elicit valid evidence of achievement.[§]

9.1.3 Separate question writing from paper production

As a supplement to this, and in order to ensure that validity is kept salient, we make one small suggestion that has wide-ranging implications:

> The two processes of writing questions and of constructing exam papers are logically separate, and should be kept apart if possible.

(Note that when we speak of ‘creating a question’ we generally mean ‘creating a question and mark scheme’ together.)

Sometimes two or three principal examiners collaborate in producing a paper, usually because they have different expertise. But it is probably generally true that different individuals differ in the kinds of questions they are best at creating, rather than being just a matter of different subject expertise. Suppose we encouraged examiners to write questions whenever they came up with a good idea, rather than just when there was a paper that needed to be produced: what difference would it make?

First, each examiner would write the questions they are good at writing, and not be obliged to write ones they find difficult to write: this alone is likely to improve question quality. Second, and more important, the question will be more likely to emerge in the way we described above, with the emphasis on eliciting valid evidence of achievement rather than to fill a gap in the syllabus specification. Although it is hard to be sure in any particular case, we saw many questions in which we think the question or the mark scheme was distorted, because some more marks were needed in one assessment outcome category and fewer in another.

When a paper is needed, the committee would meet to select questions for it. If the bank of available questions is large enough this will be very simple and will involve *only* choosing which questions to use; otherwise, some adjusting of questions and mark schemes will be necessary. But even then, this arrangement takes away the time pressure that normally attends question paper production. The key point here is that the skills involved in question writing and reviewing are *not* the same as those needed to ensure that a question paper is suitable for assessing the syllabus specification, yet the current QPEC model combines these two functions.

The principal disadvantage of this approach would be that more questions would be needed, at least in the short term, to get the system started. Like an item banking system, however, once it was operating there would be no greater demand on question production than at present.

[§] For a discussion of creating the mark scheme before the question in A Level Politics, see O’Donovan (2005).

When this suggestion has been raised with subject officers a common response has been that, despite the obvious attractions it holds, it is not feasible because of the way regulation operates. It is said that syllabus specifications are revised so often, and adherence to their quantitative terms monitored so strictly, that few questions written in one year would still be suitable three or four years later. If this is so then it should be recognised that the practice is harmful. 'Real' geography or business practice does not change so fast that what we learn one year is of no use a few years later, and good questions should be appropriate for a similar length of time. Also, syllabus specifications are meant to help examiners produce good exam papers and should not make their life more difficult.

9.1.4 Creativity

Part of the process of drafting examination questions is creative. What can Awarding Bodies do to assist examiners in being creative? There is a sizeable literature on this issue, as many organisations are concerned with creating an innovative culture. First, Awarding Bodies could try to select creative individuals as examiners, although creative thinking skills are difficult concepts to assess in selection.

Another problematic feature of the examination culture is that it tends to be a *controlling culture*, which has been argued to stifle creativity (McLean, 2005). The multi-dimensional specification that controls the content of question papers, and the weight of tradition, are very restrictive and examiners may feel that their scope for creativity is seriously constrained. As argued elsewhere, a review is needed of the constraints and their effect on the predictability of the question papers' structure and content.

What is lacking currently is a 'creativity phase' in the design of questions. Giving examiners time to be creative could be more straightforward than it sounds. Although creativity is often considered to be a feature of individuals, Shalley and Gilson (2004) review literature which suggests that creativity may result from the interactions between team members, particularly when individuals each add some form of diversity to the group. Awarding Bodies could look at the construction of examining teams and the interaction processes within the teams to try to foster creativity. Feedback from scripts and external evaluation of the teams' work may also improve creativity, although there is mixed evidence on this point (Shalley and Gilson, 2004).

9.1.5 Systematic use of Command Words

We have highlighted several times the crucial role of command words, whose function is to communicate clearly to students what it is that they are supposed to do. But this describes their role too simply, for they are part of the examination system, and they need further consideration here.

Pollitt, Walker and McAlpine (2005) found 187 different command words or phrases in one year's set of SQA exam papers. In the present study we used those as first level codes and modified the list where necessary. After some rationalisation and reorganisation, the final list contained 155 'pure' command verbs, as well as 6 interrogative adverb/pronouns, more than 20 phrasal verbs (mostly based on give, how, what and write) and two sets of questions based on the auxiliary verbs 'be' and 'do'; there were also a number of compound verbs and the rather common 'Complete the sentence/statement' question form. A complete survey of all English, Welsh and Northern Irish question papers could presumably extend this list further and we are led to ask whether it is necessary or desirable to couch exam questions in so many forms. What, for instance does enumerate mean that list does not mean? Do we really need to use all of: allocate, arrange, classify, rearrange, reorganise, and tabulate? Or order, sort, prioritise and rank?

In this study we have seen papers that make excellent use of command words. Typically they use a restricted set of question forms and command words that are closely bound to the assessment outcomes described in the syllabus specification. Given this link, we presume that teachers will ensure that students understand the message each command word carries about the nature of the task and the skills that the examiners are expecting them to demonstrate.

At A Level it is sometimes essential that students know how many marks will be allocated to, for example, knowledge/understanding or evaluation/argument, even where this is not explicit in the question paper (O'Donovan, 2005). Implicit structure of this kind happens only to a limited extent in the GCSE papers we have studied, but we think there is a greater risk that candidates will fail to understand the kind of response required by misinterpreting the command word. Since the purpose of the words in a question is nothing more than to communicate to candidates the task they are required to carry out, we therefore suggest that examiners work towards a more common and more explicit system of using command words. During the project, and especially at presentations of the preliminary findings, we have discussed ways of achieving this, but simple solutions such as a national glossary of command words face an insuperable problem in the differences in the meanings of words like 'explain' in different disciplines.

In Section 10.6 we will make some suggestions for progress in systematising command words. These are not new ideas; in one examination we saw questions which suggest that the range of command words had already been reduced: most of the command words were actually taken from the definition and description of the assessment outcomes in the specification. But it seemed that, once again, a principle for improving questions could degenerate into rather thoughtless following of rules, for some of these command words were really rather odd. This is why we suggest a number between 20 and 30 as appropriate.

9.1.6 The 'explain' problem

We have documented many problems with the command word 'explain', which is used extensively in all subjects. Its meaning is not always obvious, however. It may be used with adverbs in 'explain how' or 'explain why', or with a variety of nouns that alter its meaning such as a problem, a reason or a meaning, or as the second part of a compound command like 'suggest and explain'. In the subjects we studied it is by far the commonest command word; Appendix 2 shows that, in Business Studies, 30% of the questions used it. The second commonest was 'give *n* reasons' for something, which may be interpreted as a form of explanation as well. Thus it seems that the generic notion of 'explaining' lies at the centre of our assessment system.

Assessment of learning naturally involves students explaining their understanding of knowledge. On closer inspection, however, it turns out that explanation is a slippery concept. A definition of the word is elusive because the same language forms can count as an explanation in one context, but not in another. Shared knowledge between the speaker and the listener, and context, play a large part in determining what is needed to constitute an explanation. A fundamental problem in assessment is that the status of shared knowledge is uncertain. 'Everyday explanations' occur in conversation, when the 'recipient' can probe further if the 'explainer' assumes too much in common, or hurry them along if they are being too pedantic. In the exam explanation, on the other hand, the student can only guess at how much pedantry is required to constitute an acceptable explanation. Educational assessment is probably the only linguistic context in which a writer is required to write explanations to someone who – by definition – already knows the answer. We demand of our pupils that they correctly guess how much knowledge the expert reader will 'pretend' to share with them.

Explanation is a subtle concept. In attempting to define the word, some authors have resorted to using the word 'explain' in the definition itself (eg *Oxford English Dictionary*; Wiener, 1989). Chandrasekeran (1988) wrote that explanation means 'To make something clear, to make the listener understand'; by this criterion exam answers can only be quasi-explanations, since the examiner understands already.

A typology of explanations could not be found in the literature. Philosophers disagree about whether explanations are essentially deductive (Hempel, 1965), statistical (Salmon, 1970), pragmatic (van Fraassen, 1980), or linguistic (Achinstein, 1983). Different subject domains treat the notion of explanation differently. In psychology, for example, a mathematical equation is no longer considered an explanation for psychological phenomena, but in mathematics, the working through of an equation may be just the kind of explanation that the examiner is looking for. Factual statements and equations may count as explanations in chemistry, but in history the examiner may expect more information about the temporal and political relations between events if one is to explain the other. For some subjects, prediction from theory is important to a good explanation, whereas in others, explanations are more akin to descriptions of events and the relations between them, with expectation of prediction.

It is worth remembering that GCSE pupils are not young mathematicians or chemists or historians: they are 15-16 year old children, with just one mind to study as many as a dozen different disciplines and to master the range of meanings that ‘explain’ may have across them.

9.1.7 Cognitive levels

While Benjamin Bloom’s taxonomy of educational objectives is well known (Bloom et al, 1956), its application to UK exams has been problematic, as the six levels he describes do not map easily onto our typical questions. E. A Peel (1971) and his students in Birmingham systematically explored the understanding of pupils and developed a classification of their responses that should be more useful. He described three principal levels of response to questions asking for an explanation of ideas in a text presented to the pupils:

Mentioning – tautological, partial or inconsistent response
Describing – a mainly correct list of aspects of the passage
Explaining – responses which used additional related ideas to interpret the meaning of the text

In addition he described a more developed form of explanation involving “a capacity to combine more than one piece of evidence with outside ideas to evoke cause and effect.

From this, and from a neo-Piagetian perspective, Biggs and Collis (1982) developed their SOLO taxonomy which uses five levels labelled: Prestructural, Unistructural, Multistructural, Relational, and Extended Abstract. The first two correspond roughly to Peel’s ‘mentioning’ category, and the other three interpret his higher categories.

We suggest that Peel’s system is both simple and useful for GCSE examinations, and would gloss the four levels in his extended version as:

Mentioning – in which concepts are named
Describing – in which attributes are identified, listed and described
Explaining – in which meaning is expressed by relating concepts causally
Judging – in which alternative meanings are evaluated

The fourth level can be split further into:

Justifying – in which the theoretical basis for a proposal is explained
Evaluating – in which both sides of an issue are assessed and a balance judgement made

We see clear evidence of this kind of scheme in the definitions of Assessment Outcomes in some syllabuses, and in the choice of command words in at least one series of exam papers in this study. Such a scheme could be used to define four sub-sets of command words that an examination will limit itself to, could be conveyed to students to make clear what kind of evidence of understanding is expected in different questions, and used to assess rapidly (and roughly) the overall level of cognitive processing demands in a paper. This last, however, would require that the command words be used consistently.

It was mentioned above that ‘give *n* reasons’ can be seen as a kind of ‘explain’. There is, however, a crucial difference. The Semi-Constraining request to ‘give two reasons’ why, for example, a town is located where it is will be satisfied if a pupil gives *any* two from a potentially long list of contributory factors. In contrast, the Un-Constraining request to ‘explain’ why it is there will only be satisfied if the reasons the pupil gives are the most important reasons in the list. In the Ross-on-Wye example earlier we remarked that a pupil could score full marks without mentioning the river, since the command ‘Add detailed labels … to explain’ amounted to asking for reasons. But had the question been Un-Structured, an answer that did not mention the river would have been rated as poorer than one that did. In Peel’s terms, ‘give *n* reasons’ is on the border between *Describe* and *Explain*.

9.1.8 Understanding expectations

Understanding what is demanded by the various command words is one part of a larger issue about expectations and predictability.

At age 16 school pupils are still generalists, studying as many as eight or ten subjects. They will have developed a rather general *schema* for exams, that is a set of expectations about what will happen in an exam – where and when it will take place, what people will be there, how they will feel before and during it, what they will experience and what they are supposed to do. For each separate exam they will add to this some specific detail, especially concerning the knowledge and skills they expect to be asked to demonstrate and their level of confidence in that subject. A critical part of this schema is the pupils' expectations of how to answer exam questions, which they will have developed from practice, with the help of their teachers and textbooks (Crisp, et al, in press).

In many ways students' expectations and their exam schemas can help examining. For example, students will respond in particular ways to particular command words, and the overall structure of the examination will be familiar. But too much predictability can harm the process, by allowing candidates to gain marks by reproducing memorised answers without thinking too hard about the demands of a particular task. This, of course, is one of the reasons why so many questions are set in context, requiring knowledge to be applied to the specific case rather than just repeated as in the textbook.

The issue of the predictability of examinations is inextricably linked with students' expectations. Is predictability a good or a bad thing? In general, the essence of every examination task *should not* be predictable; that is, the exact subject content, and the skills and cognitive processes required.

What *should* be predictable is anything *peripheral* to the job of getting students to show us what they know and can do. This includes the formats of questions and responses, vocabulary, grammar, how resources like maps and diagrams are to be used, and anything else that is involved in interpreting the question. Also, the criteria that will be used to assess the responses should be predictable.

It is very important for examiners to understand students' expectations, and to strike an appropriate balance between predictability and demands.

9.2 A taxonomy of mark schemes

In Step 3 of our ideal model for writing questions we state that the right type of mark scheme should be used to capture the outcome space that is to be elicited. During this study we developed a classification of mark schemes that recognised 22 different sorts of scheme, including a type with “No guidance”. The other 21 were classified in three sets. That scheme was more elaborate than necessary, and some of the distinctions bear no theoretical importance. We now recommend that examiners should consider the simplified typology of mark schemes below. It is divided into the same three sets that we introduced at the beginning of this report to describe the degree of structural constraint a question gives to, or imposes on, a student’s answer or, equivalently, on the range of the expected outcome space.

In general, the quality of the mark scheme improves as we go down the sub-types within each set, although the ‘lesser’ types will certainly be adequate for some questions. Note that it is the response that is or is not constrained rather than the stimulus, but we are using the word ‘question’ here to stand for all of the question, the answer and the marking. Set 3 involve most constraint while Set 1 involve least.

Set 1 Un-Constrained questions

UC . 0 Model answer

A perfect answer to the question – in the opinion of the question setter.

From the evidence in this study it seems that the day when a Principal Examiner wrote a ‘model answer’ and offered it to other examiners to use as a mark scheme has, at last, ended. We still detect vestiges of the old approach, however, and we have discussed the problems with it earlier. The key point is that it does not help markers deal with real answers because it is too good to serve as a useful comparator. Consider this example:

Example :

EPP offers a range of products that appeal to different age groups.
ii) Why does it do this?

MS:

Valid points could include:
20 m customers in the market – appeal to different segments
increasing products offered – spreads risks/increases sales
cater for different customer tastes – expand market share
all designed to increase sales/secure future/expand

Level 3
Candidate makes sound judgements 6-8 marks

The UK cinema market is over 20m and EPP want to increase their share of this market which they do by appealing to different age groups. Increasing the different products they sell will appeal to more people eg tshirts, gift vouchers, monthly regular visitor cinema passes. Different people have different tastes and therefore need different things – retired people might prefer to go to the cinema through the day so EPP have special viewings and events for them. All of these strategies are designed to increase their market share, compete with other cinemas and expand.

Level 2
Candidate makes basic judgement 3-5 marks

Level 1
Candidate applies knowledge 1-2 marks

The illustration of a Level 3 answer here is so good that it really operates like an old-fashioned model answer; the danger is that markers will rarely award a Level 3 mark to real answers because they will not measure up to this standard of quality and completeness.

UC . 1 Holistic implicit levels

Rating against a set of holistic level descriptors.

In general, for unconstrained questions model answers have been replaced by sets of level descriptors. Our type UC.1 consist of schemes of the kind described as ‘holistic’ in the language testing world, and as ‘best fit’ in other contexts. Here is an example from the present study:

Example UC . 1 :

MS:

(d) CASE STUDY:
Population change and how it affects a place. Not just numbers but groups.
Could be a LEDC or MEDC city. L2 max if inappropriate example or candidate uses
Hong Kong. Generic case studies = Max L2
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]

The descriptors are very generalised (‘generic’) and include several qualitative traits. Problems arise if these traits are not highly correlated, since a particular response may rate significantly higher on one trait than on another. UC.1 is appropriate for assessing *pupils*, as in national curriculum assessment, when a given child must be ‘in’ one of the levels, but it is not generally appropriate for assessing extended responses in content-based exams. With many ‘real’ responses, a marker will feel that the pupil deserves different levels for different traits, but the mark scheme gives them no help in choosing the right overall score.

UC . 2 Multiple levels - implicit or explicit

Rating against a set of analytic rating scales

To deal with the problem of confounded traits, some mark schemes divide the descriptors of the main traits into separate columns. In language testing this is called ‘analytic’ scoring; while analytic assessment of skills-based traits like ‘speaking’ may involve around five subscales, there are rarely more than two in GCSE content-based examinations.

We can identify two versions of type UC.2, depending on whether or not raters are given explicit instructions on the relative weighting to give each subscale. An example of each is given below.

Example UC . 2 – implicit :

Explain the formation of a corrie and its lake.
Add labels to the diagrams below to illustrate your answer. (6)

MS:

Level	Explanation	Communication
Level 1 1-2 marks	Some understanding of the nature of corries. No or irrelevant labelling.	Information is communicated by brief statements.

Level 2 3-4 marks	Understanding of role of glaciers and can name processes. Basic labelling. NO LABELLING = LEVEL 2	Communication may be verbose or illogical. A limited number of specialist terms are used. There is some accuracy in spelling, punctuation and grammar.
Level 3 5-6 marks	Clear explanation of abrasion and plucking / role of deposition / tarn formation. Clear and well labelled.	The written style has a suitable structure. There is a range of specialist terms. Spelling, punctuation and grammar have considerable accuracy.

This levels mark scheme has two scales, one for the explanation and one for how it is communicated. There is no indication of how these two scales should be combined to give a final mark.

Example UC . 2 – explicit :

(b) Discuss whether penetration pricing would be the most appropriate pricing strategy for CP to use for its new conference business customers. **(8)**

MS:

Possible areas for discussion:

For:
may kick-start interest from new business customers
Price can creep up over time without losing business

Against:
loss of potential revenue
retaliation from competitors, eg price war
equating price with quality

	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)

In this case, the relative weight of the two scales is made explicit by scores for each level *within* the two traits.

Again, these rating scales tend to be generic, with little difference when they are applied to different questions. This may sometimes be a problem if the students' interpretation of the importance of each trait varies according to the particular task.

UC . 3 Specific trait interpretation

The general trait as it is realised in this particular task.

Types UC.1 and UC.2 are skill-focused, in that they aim to assess very general cognitive or behavioural skills. In contrast, type UC.3 is task-focused; there is just a single scale which aims to differentiate performances simply in terms of how well they perform on the task.

In responding to a 'question' students are actually trying to perform as well as they can, in the belief that a 'better' performance will gain them more marks. In a well-designed exam task this kind of good performance will be a direct indicator of how well the student can 'do' the subject as a whole; in a good UC.3 mark scheme the assessment scale will directly represent the overall trait of achievement in the subject. If the mark scheme links the scale for a question to the overall trait markers should be able to decide for themselves how important the several component skills are in a particular task. In language testing this task-focused approach is implemented under several names, including 'primary trait scoring' or 'communicative effectiveness'.

The two examples below show it in use in Geography and Business Studies:

Example UC . 3 :

(v) Cliff recession causes many problems for people who live in coastal areas.
Choose a case study of a stretch of coastline or coastal area that is suffering from cliff recession.
Chosen stretch of coastline or coastal area. ……………………………………….
Explain the causes and effects of cliff recession in this area.

MS:

Levels mark
Walton case study or other.
Section 1A of appendix
Do not credit management

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.

(5)

Here the overall trait is interpreted as 'Able to explain particular instances of cliff recession' and interpreted in three levels: Description; Incomplete Explanation; Complete Explanation.

Example UC . 3 :

(e) Suggest and evaluate ways in which Emma can expand her business. [7]

MS:

Level 1	[1-2]	Suggests way(s) in which the business might expand but with no evaluation.
Level 2	[3-4]	Suggests way(s) in which the business might expand with one - sided or unsophisticated evaluation.
Level 3	[5-7]	Suggests way(s) in which the business might expand with well-balanced/sophisticated evaluation.

Suggestions might include:

- merger;
- takeover;
- diversification;
- internal growth;
- labour specialisation;
- franchising – may consider advantages and disadvantages;
- elements of marketing mix, e.g. selling the service in a wider area.
- etc.

General advantages might include:

- economies of scale;
- greater sales/profits;
- etc.

General disadvantages might include:

- diseconomies of scale;
- loss of control of business;
- communication;
- cost of growth;
- etc.

Some candidates may interpret this question as the method by which business expand and this must be fully credited. Others may look at it in terms of raising finance for expansion again this must be fully credited.

In all cases in Set 1 the focus should be on assessing the quality of the response, but markers will be helped by indications of what kind of content is appropriate and how to deal with inappropriate content. The student's ability to choose appropriately has to be included as an element in the overall judgement of the quality of their response.

Set 2 Semi-Constrained questions

SC . 0 No guidance, model answer

A circular statement that adds nothing to the question

Just as with the Un-Constrained questions so with Semi-Constrained ones we can find model answers and other unhelpful mark schemes.

Example SC . 0 :

(i) What is a quality circle?

MS:

(i) a group of employees that meets to identify quality problems, thinks of solutions and makes recommendations 1

Many pupils will give an answer with *some* of these points in it, but markers are not told how much is needed to earn the single mark.

Example SC . 0:

(c) The clock shown in Fig.3 must be able to stand on a table as shown below.
Use sketches and notes to show how the clock could be made to stand on its own.
Include details of materials, fittings and methods of construction used.

MS:

(c)	Some form of stand/support	0-2	
	Appropriately named materials	1	
	Fittings and/or constructions used	0-2	**[5]**

Here the mark scheme does help with the allocation of the marks into three sub-parts. Apart from this, however, it hardly adds anything to the question, merely restating the requirements. Markers are not helped in deciding whether a particular response is appropriate enough, or good enough to get credit. What would not count as “Some form of stand/support”? This is a common problem when the response is drawing rather than writing, and we do not underestimate the difficulty of describing in words the criteria for assessing sketches.

Perhaps it might be better to assess these questions more for quality than for ‘points’; otherwise the mark scheme must try to define the nature of acceptable ‘points’ in the diagram.

SC . 1 List of Good responses (Examples or Complete)

A listing of Outcome Space **Good 1,2**

This is the traditional ‘points’ mark scheme, where every possible creditworthy point is indicated. If it is complete it probably maximises reliability, so long as all markers are prepared to abide by it. Type SC.1 is sometimes adequate: when only one mark is to be awarded for any relevant point, or when several marks are available and every point in the list is equally important. It is essentially a counting system with almost no judgement involved and represents an extension of the objective marking system from Very Constrained questions into the realm of less constrained ones. It is not appropriate when some points are better or more important than others; it fails to make any distinction between content and quality, and may leave markers unsure whether to award a mark for a mere mention of a point with no evidence that it is understood.

It is, of course, difficult to be sure in advance that any list is complete: consider the following example:

Example SC . 1 :

(d) Complete the table to show **three** main ingredients used to make the *filling* for your product.
Give a **different** reason why each ingredient is used.

MS:

Any of the 3 main ingredients used in the FILLING of the chosen design 3 x 1 mark
Relevant reasons given for use of main ingredients e.g.

- binding
- adding texture
- adding flavour
- adding nutritive value
- adding colour
- moisture
- aroma
- enriches 3x1mark
- adds bulk / main ingredient / structure
- accept generic names of ingredients e.g. meat, vegetables (6 marks)

It is common for SC.1 mark schemes to introduce the list with ‘e.g.’, or to complete it with a phrase like ‘Any other relevant answer’ – even when, as here, it looks as if it was intended to be a complete list. Note that no guidance is given to judges on how to decide if an unexpected response is ‘relevant’ enough to get credit.

SC . 2 *List of Good + list of Poor (Examples or Complete)*

A listing of Outcome Spaces **Good 1,2** and **Poor 1,2**

As with Very Constrained questions a ‘points’ type mark scheme can generally be improved by taking care to specify things that candidates are likely to write that are *not* good enough to merit a mark. One reason why this is helpful is that it necessarily leads the question writer towards considering the quality of the students’ responses and away from thinking only of counting how many relevant things they mention.

Example SC . 2 :

(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

This mark scheme tries to define both the Good and the Poor outcome spaces. Both lists seem just to give examples, with no attempt to be complete but there may, in fact, be no other answer that students are likely to offer. In any case, considering the difference between ‘good’ and ‘poor’ is helpful. The list of **Poor 1,2** may become quite extensive:

Example SC . 2 :

(d) Name **one** other plastic material suitable for making the salad servers.

.... (1)

MS:

(d) Named thermoplastic
e.g. PVC / ABS / polythene / HD / LD / HIP / Nylon / PET / Polypropylene (pp)
Do not accept:
Polyester / generic term thermoset / thermoplastic. (1)

A serious attempt to think of all the responses that students are likely to offer – the Outcome Space – and how to classify each of them as Good or Poor, is likely to lead the question writer to the next and generally better type.

SC . 3 Rule/principle to differentiate responses

Defining the difference between Good and Poor

When you try to write rules for deciding whether a response is Good or Poor you find that you are describing the construct of achievement in the subject that you are trying to measure. The distinctions that you make between better and worse answers describe the differences between higher and lower grades, and almost guarantee that the questions will make a significant contribution to assessing that construct validly.

We will discuss this point further in Section 9.3. For the moment, note that it applies equally to deciding whether a single point is made well enough to merit one mark, or to deciding whether a more complex response is good enough to merit three rather than two marks. In all of the examples of SC.3 here, the examiners have tried to define the difference between Good and Poor, rather than leaving it to the markers' judgement.

Example SC . 3 :

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

MS:

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

The key idea is that renewable in this context means that it will not run out. It is taken for granted that responses will refer to sources of energy in an appropriate way (this could have been exemplified but it is here assumed that all markers will know them well enough). The phrase about "Bottom line" is used to define the boundary between the outcome spaces Good 1,2 and Poor 1,2.

Example SC . 3 :

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)

The first sentence in the mark scheme explains to the marker what the question writer was aiming for in this question. As with type UC.3 the essential point here is that the mark scheme translates the overall trait of achievement into a form specific to this task.

The 'deforestation' examples discussed in Section 8.2.7 illustrate neatly how a question can be improved by moving 'up' the scale of UC marking types. In the first version we have just a model answer:

Example SC . 0 :

(i) What is the meaning of the term 'deforestation'? (1)

MS:

Removal of trees

In the second this is improved by adding other acceptable answers:

Example SC . 1 :

(i) What is the meaning of the term 'deforestation'? (1)

MS:

Removal of trees, cut down, burning, destroyed.

In the third version, responses are classified in terms of their quality on the general trait as realised in this task, yielding more evidence about the students' level of achievement:

Example SC . 3 :

State the meaning of the term **deforestation**. (2)

MS:

Level 1 ([1])

An incomplete definition

Cutting down a tree

Removing trees

Level 2 ([2])

A full definition needs some reference to scale

The complete clearance of a forest area by cutting down or burning trees

Once again, we cannot expect a mark scheme on its own to deal with the whole of the potential outcome space. If there are many unanticipated answers and markers are finding it difficult to decide on them using the rules in the mark scheme, and if expected answers are not appearing, then the fault lies either in the question or in the question writer's prediction of how students would respond.

Set 3 Very Constrained questions

With these, the candidate has almost no freedom in responding, as the precise format of response is given. There is virtually no response demand, and assessment is almost entirely concerned with correctness.

VC . 0 No guidance, model answer

One acceptable response out of several, or a circular statement

Mark schemes of this type tend to make assumptions about how students will respond that are unsafe, often based on the expectations of experts rather than learners.

Example VC . 0 :

g) Complete the following sentence about private sector businesses:

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

……………………………………………………………………………………………… [1]

MS:

(g) The capital of a private business is contributed by **the owners/shareholders.**

These two phrases may be the *best* ways to complete the sentence, but what should a marker do about "investors", "the people who set it up", "capitalists", "banks", "financiers", etc? Are any of these close enough to deserve a mark? The examiners intended this to be a kind of vocabulary test, seeking the 'right term' for each definition, but there is no constraint on candidates to treat it this way.

Another common use of model answers is in calculations where, given the programmatic nature of a standard calculation, it may suffice. But not always:

Example VC . 0 :

(d) Calculate the total weekly wage bill for Carl's Cars for the eight mechanics each employed for forty hours at £11 per hour. (Show your workings.) [2]

MS:

Wage bill	=	440 × 8	(1)
	=	3520	(1)

The problem comes when there is an alternative way of working out the answer. Suppose a candidate tries to calculate the total numbers of hours first – and makes an error:

Wage bill	=	340 × 11	(1)
	=	3740	(1)

The mark scheme gives no help in a case like this.

VC.1 Complete list of Right answers

A complete list of acceptable answers

In type 3.1 the Outcome Space **Good 1** is fully described. A trivial example is a multiple choice question where **Good 1** is fully described by declaring that the correct answer is "b". In other cases a complete list of all possible acceptable answers is given.

Example VC . 1 :

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 (1)

In this case, it is quite easy to give markers a 'complete' list of good answers: it's hard to imagine any other acceptable response that differs significantly from these.

The 'hardwood' example from Section 8.2.4 shows how this type may fail:

Example VC . 1 :

(i) Name a hardwood commonly used in the manufacture of children's toys. [1]

MS:

Hardwood: beech. [1]

In the mark scheme, "Hardwood" is a type VC.0 scheme, since it does nothing more than restate the question; "beech" is type VC.1, apparently the only acceptable type of wood. As discussed earlier, however, alternative right answers – outcome space Good 3 – can be identified.

VC.2 List of Right + list of Wrong (examples or complete)

A listing of Outcome Spaces **Good 1,2** and **Poor 1,2**

As with Semi-Constrained questions, it will generally help markers if examples of unacceptable responses (outcome space **Poor 1,2**) are given. When they meet a response that is not quite as good as one that is listed as acceptable markers are now helped decide if it is good enough:

Example VC . 2 :

(ii) Name **two** other permanent joints which could be used for the corner joint of the wooden frame.

1. ....

2. .... (2)

MS:

c(ii) Mitre/dovetail/comb/finger/glue and screw/glue
and nail/glue and pin/rebate/butt/biscuit/dowel 2 x 1 (2)
(Do not accept KD fittings/mortice and tenon)
(do not accept just glue / screw / nail / pin)

The transition from VC.2 to VC.3 is similar to that observed with SC question types: when an examiner tries to produce a complete list of acceptable answers together with an indicative list of wrong ones it often becomes clear that it would be easier to define the boundary between good enough and not good enough.

VC.3 Rule/principle to differentiate Right from Wrong

Defining the difference between Right and Wrong

In principle, type VC.3 is better than VC.2, since a principle or rule has been laid down which markers will be able to use even with answers that the original examiners did not anticipate. Thus this type copes with outcome spaces **Good 3** and **Poor 3**, the unexpected range or responses that a marker may meet.

A simple case of VC.3 occurs with estimation questions such as this graph-reading one:

Example VC . 3 :

(ii) How many garments does Badge Identity Ltd need to sell in order to breakeven?

MS:

3000 (allow 3000-3200) I

The “allow” statement defines exactly the whole of OS Good 1,2 and, by implication, the whole of OS Poor 1,2 as well. If there are no other formats that should be expected (apart from trivial ones like *3,000* or *three thousand*) this is sufficient to deal with all possible responses.

Typology conclusions

In our analyses we used a more complicated system for describing the mark schemes we saw: our aim there was to be *descriptive*. In this section we have laid out a fairly simple scheme that captures the extent to which a mark scheme is able to help guarantee validity, understood as a close mapping of the trait of achievement we want to assess in the exam onto the scale that actually rewards students for the quality of their responses to an individual task. In that sense this scheme is *prescriptive*; in general, a higher level of mark scheme within each of the three question types will be better than a lower one.

Nevertheless, as we have noted several times, a lower rated mark scheme may sometimes be perfectly adequate; type VC.1 will always suffice for a multiple choice item, for example. It will always be the responsibility of the question and mark scheme writers to judge when a scheme is going to be good enough.

The typology is not a complete specification for how to write good mark schemes, but the basic classification seems to us to capture the most important principles. Other criteria should not be ignored, such as whether to choose a UC or SC type for a given task, how to specify scoring rules, or when and how to blend different mark scheme types. We hope to develop advice along these lines further in the future.

9.3 Ensuring Construct-relevant assessment

Several of the points made in these Conclusions and Recommendations may be understood as aspects of what is probably the most general principle we can express to improve quality assurance in examining. Underlying all good assessment will be a constant concern that every question we write, every differentiation we make, every decision to award or not award a mark should be made in accordance with the construct we are trying to measure.

In giving one student a B grade when another gets a C we are declaring that the first one is better than the second in terms of achievement in this subject, on the basis of the evidence we saw in this exam. We are saying that the first one has achieved more than the second, that they have ‘more of the construct’ we have measured. This relative judgement will only be accurate and fair if we have ensured that all of the decisions in the test were in line with what we mean by having more ability in Geography or in Design & Technology. ‘More ability’ is interpreted in practice to mean showing more knowledge of facts and principles, or being able to give clearer or more exact descriptions and explanations, or being able to deal with more difficult or more complex tasks, or to make a case more persuasively – all of these constitute evidence we would trust in making our grading decisions. Note that repetition, in which a student does the *same thing* more than once does not amount to evidence of a higher level of achievement: little is gained by asking students to do the same thing several times in one examination.

There are other kinds of evidence that tell us about the student’s abilities but which are not relevant to assessing the trait the whole exam seeks to measure. Over the years we have deliberately changed the demands we make, for example, on the ability to memorise poems, textbooks or lists of geographical features; we have also reduced the demand on the ability to ‘second guess’ the examiner by making the questions more transparent. We have changed the amount of reward we give to students for the ability to write well-crafted essays, or to carry out lengthy calculations, or to compose a convincing argument. An examining team must be clear, before they create questions, just how important each of these demands is and must design their tasks accordingly.

It is for these reasons that we insist on the importance of:

1 being clear about the *key idea* of a question;
2 careful consideration of the *intended outcome space*, the range of evidence we want to see;
3 careful design of a mark scheme that will evaluate that evidence in terms of the overall trait the exam aims to assess;
4 care in wording the question to ensure that we will elicit that evidence.

Beginning with our principle for validity in assessment, we argue that if the students' minds are *not* doing the things we want them to show us they can do then any judgements we make of their resulting performances cannot accurately reflect differences in the amount of construct-relevant ability they have. We have documented examples where questions have failed to ensure that their minds are properly engaged in this way.

Then we addressed inadequacies in the mark schemes being used to evaluate these performances. Again, if students are not rewarded properly for doing the things we want to see that they can do then their relative scores will not be trustworthy indicators of their relative construct-relevant abilities.

In the last thirty years there has been substantial improvement in the writing of questions, brought about through the efforts of many people and encouraged by the increasing openness of the system. Mark schemes have also improved, since even twenty-five years ago they were treated as 'strictly confidential'. We feel that there is still considerable scope for improvement in the systematising of mark schemes and, in particular, in matching the type of mark scheme better to the type of question to make sure that the evaluation of students' responses is always made with the appropriate aspects of the achievement construct firmly in the examiner's mind.

9.3.1 Why writing questions is difficult in the subjects of this study

It is not hard for new examiners to find general advice, and a great deal of experience and practice, on which to draw when they first try to write questions. It is, however, more difficult to write questions in subjects like Design & Technology or Business Studies than in more traditional subjects, simply because they are relatively new subjects with less history of assessment to draw on.

But there is more to it than this. Over a long period an academic discipline develops its own traditions, its own body of knowledge becomes more separate from 'general knowledge', and its technical terminology becomes more separate from everyday language. Few people, for example, would confuse the meaning of 'resistance' in a physics exam with its meaning in everyday life or politics, or the meaning of 'product' in a chemistry exam with its meaning in everyday shopping (although, alas, the 'few' include some anxious and uncertain GCSE science students). But politics and business and technology are seen as part of everyday life in a way that physics and chemistry are not, and their vocabulary is not so discrete. It is not so obvious what can or cannot be called a 'cost' or a 'factor' in business, or a 'finish' in food technology. This is presumably one reason why there are so many questions of the type 'Explain what is meant by the term X' in these subjects.

Psychologists have shown that there are two distinct cognitive systems that we use for solving problems (eg Evans, 1989). One is fast, instinctive and approximate, based on experience of how the world usually works, and is rooted deeply in our evolutionary history. The other is slower, more exact and certain, based on logic and reasoning, and has to be taught and learned in school. It is, of course, the second kind of problem-solving that we seek to assess in exams *and not the first*. To the extent that a subject has not yet become separate from the everyday world, assessment will be difficult.

This 'real world' problem is exacerbated by the use of real world contexts in examining. A facile response would be to recommend abandoning contextualisation from exams, but we are convinced that there are good reasons not to do this; in particular, the ability to apply logical reasoning rather than to go for the 'obvious' explanation or to the 'implied' conclusion is one of the marks of developing expertise. The constant risk that everyday reasoning may intrude ("pre-cognitive bias" in Evans' scheme) simply means that question writers must be more vigilant and devise more procedures to ensure that their students' minds are indeed not misled before they can start doing the things we want them to be doing.

10 Recommendations

10.1 OSCA Training

The system we have proposed – Outcome Space Control for Assessment – has evolved out of several years of research and from our experience in observing and training examiners. It is a systematisation of the best practices we have observed, underpinned by theoretical principles based on our own research and that of others. In this project we have been able, for the first time, to review the practice of all five awarding bodies in England, Wales and Northern Ireland, and to present our conclusions to a representative sample of examiners. As a result we are confident that, given suitable training, most examiners would be both able and willing to follow the OSCA system, and that they would find it helpful in raising the quality of their GCSE exams. **Our first recommendation is therefore that examiners should be offered training in the principles and procedures of the OSCA system.**

But while many of the procedural aspects of OSCA are already accepted as good practice, the rationale for them is very rarely explicit. Senior examiners spoke to us of having learned to write assessments by trial and error. Normally, they are appointed to these senior positions as a result of being a successful marker and of managing a marking team well, but the skills needed to mark consistently and to train a team of others to mark consistently are very different from those required to write the assessment materials themselves. **As a priority, more education and training in the fundamental principles involved in creating questions and mark schemes should be provided for current senior examiners.**

The examiners involved in the QCA Examiner Conference agreed that this kind of training is needed by everyone involved in the question paper production process, that is the Reviser and the Scrutineer as well as the QPEC members. It was also suggested, and we agree, that **awarding bodies should consider the feasibility of adding a language reviewer to the team**, whose responsibility would be to ensure that the language and presentation of the question would succeed in conveying clearly to all candidates the nature of the task that the examiner wanted them to do.

10.2 Question level statistics

Training is only the start of professional development; examiners need to see the positive effects of using the system in questions that actually function better. To enable this, **awarding bodies need to make more efforts to feed back to their question writing teams evidence about the functioning of their questions**.

This will become much easier with the advent of on-screen marking, and the consequent easy availability of question level data. A simple routine item analysis would suffice to show examiners how successful their questions were. There are two ways in which a question can fail to function as intended: discrimination or misfit statistics will show when it fails to measure the same trait as the rest of the exam; the average score will indicate when a question turned out to be unexpectedly easy or difficult. If these statistics can be fed back to the examiners soon after the exam took place, they should usually be able to explain, from their experience of marking responses, why the question failed to function as intended.

For constrained and semi-constrained questions, unexpected difficulty or easiness is the most powerful indicator of success. Examiners generally target these questions at particular grades – F/G, perhaps, or A/B/C. If the average score turns out to be much lower or higher than is appropriate for these grades, then either the question has failed to elicit the intended outcome space or the mark scheme has failed to evaluate it properly. We therefore suggest that, as part of the question writing process, **examiners should predict the average score the students will achieve on each question**.

The best way to ensure that questions will function as intended is, of course, to pre-test them before the live examination. It is unlikely, for various reasons, that this will become common practice in Britain and examiners must seek other ways of incorporating quality. Rapid statistical feedback, and a culture in which examiners learn from their successes and mistakes, offer an opportunity for quality improvement that has not been feasible until now.

10.3 Detrimental features of the current procedure

A seminar and a conference were used to present OSCA to senior examiners and awarding bodies, and to collect feedback. The proposals were generally well received, with several examiners commenting that good practice in their boards is rather like what we were proposing. Yet our report shows plenty of evidence of bad practice as well as good: the good practice needs to be systematised, which is what OSCA does, and needs to be disseminated more widely through training so that all examiners follow it routinely.

One problem which may be causing low standards in question quality is the emphasis that seems to be placed on the process of paper production rather than on the product. High quality content in assessment materials has a very low priority in the system. Instead, the focus is on meeting deadlines, avoiding errors and following the requirements of QCA's Code of Practice and Awarding Bodies' own procedural guidance. This is all very time-consuming, leaving little time for considering how well the question and mark scheme will function as an assessment tool.

Reviewing the questions in a paper is largely the responsibility of subject experts, who often work in senior posts elsewhere in education; for them, examining is a part-time activity, second to their main life and career interests. The system itself provides little incentive for them to write a high quality assessment: it could be argued that several features of the system foster mediocrity.

Much of the content of question papers is prescribed. First, the questions must relate to the subject content as prescribed by the syllabus ('specification'). Second, the assessment objectives provided by QCA must be complied with; these relate to knowledge, application of knowledge, evaluation, etc. and carry weightings which the senior examiner must ensure that her question paper abides by. Third, it is the custom for Awarding Bodies to produce assessments that do not surprise teachers and students in their style or structure. Thus, if there were five multiple choice questions at the beginning of the question paper last year, there are likely to be five multiple choice questions there again this year – unless teachers have been clearly and repeatedly warned otherwise. Schedules for question paper production typically range over two years; during that period, the question paper is revisited several times for review, amendment and checking by the senior examining team and Awarding Body staff. With this level of prescription, the committee method of handling the content, the pressure of deadlines and the lack of time, the system encourages a handle-turning approach to question paper production: it is a prescriptive and laborious task. The syllabus specification, the specimen assessment materials and the system now control the process they were originally meant to support.

Three kinds of skill are needed to create good exam papers: managerial, academic and assessment skills. The current systems, with their emphasis on processes, prioritise the managerial dimension at the expense of the others. The work is largely carried out by people originally chosen for their academic expertise, who need to develop the managerial skills to do it. The third dimension, the skills needed for competence in the professional task of assessment, is largely ignored and, within it, the skills of question creation are almost unknown. **We recommend that new question creation systems are devised that ensure the participants are able and encouraged to develop both aspects of professional assessment competence.**

10.4 Separate question writing from paper production

These three skill sets underlie our argument, in Section 9.1, for a separation between the two processes of question writing and paper production; question writing makes demands mainly on skills that are largely irrelevant to the process of constructing a paper to meet institutional and specificational requirements.

It is not necessary to change to a complete item banking strategy to gain the advantages we envisage; they arise just from recognising the difference in the skill sets demanded, and reshaping the system to be more flexible. **Awarding bodies should devise schemes that encourage examiners to create questions before they are needed for a particular paper**; **regulators should seek ways of making specifications more flexible so as to extend the potential use of intrinsically good questions**.

10.5 Get rid of bad habits

In our analyses of current question and mark schemes we saw evidence of earlier initiatives concerned with the quality of assessment. Some lessons have been learned and many of the problems seen in Pollitt et al (1975) have largely disappeared. But we also saw evidence that good advice sometimes degenerates into bad habits.

We want to encourage examiners to think about the wording of their questions and to understand how students read an exam question. One important point is to use highlighting – bold, italics, capitalisation or underlining – in accordance with OSCA theory. This means that there should be no other constraints on how tools such as highlighting are used. Highlighting is intended to emphasise an element of a question, and it should be used expressly to help students understand the task they are being asked to carry out.

Similarly, guidance for examiners to use a particular word or phrase, or to highlight certain words in a question just because it is 'house style' is contrary to our recommendations. There should always be a substantive justification for any such decision, in terms of how this particular question will be less confusing for these students.

We noted quite widespread use of gap-fill questions, particularly in Geography and Business Studies in the Foundation tiers. Especially when multiple choice options are given for students to fill in the blanks in a sentence, the question ends up testing grammatical knowledge of English rather than the subject intended. If the student’s English was adequate, the question often degenerated into a 2-option item supposed to test knowledge of Geography or Business Studies.

The only place for cloze tests is in language testing.

10.6 Command words

At the seminar and conference it was generally agreed that the use of command words merited attention, but it was less clear what should be done. In Sections 8 and 9 we have illustrated some of the problems. Many of them arise from a mismatch between the apparent meaning of the command words used and the actual demands of the question but others come, more fundamentally, from the nature of language and education.

Children do not learn the meaning of these verbs from dictionary definitions, but from the experience of meeting and using them in contexts. As with all their academic schemas, these meanings are rather undifferentiated in the early years of schooling and the children are guided more by what they are used to doing in tasks like this than by the actual command word. As they learn to partition their work into different disciplines so they begin to distinguish the commands that are more common in each, and the command words acquire meaning for each pupil by this process. Thus each word – explain, justify, describe – acquires a slightly different meaning in each discipline. For adult teacher experts this is not a difficulty, as they recognise the subject setting for a question and know what it is usual to ask about it but GCSE pupils are neither adult not expert in any single subject – they are children in the middle of a process of choosing which areas, if any, to specialise in.

There are therefore no ‘true’ meanings for many of the most commonly used command words (as we discussed for ‘explain’ in Section 9.1.6). At best, it might be possible to define a command word as it is normally used in GCSE in one subject, though even then we might find significant differences in its use in different syllabuses or by different awarding bodies. The language of a discipline – its argot or jargon – is part of what a student needs to learn as part of the process of initiation into the discipline, but it is unreasonable to expect much of this at GCSE. The problem of learning the meaning of ‘technical terms’ is well recognised but the issue of the particular flavour a non-technical word like force, circle, value – or explain – takes on in a particular subject is far more subtle and difficult to control.

We do not think that the answer to the command word problems is a glossary. It is not possible to convey the significance of a particular command word or phrase by describing it in terms of other ones, just as this is not how its meaning develops. An encyclopaedia would be more appropriate, with the many uses of each term being displayed through examples from the various subjects in which it is used. This would, however, be an enormous task, if attempted simultaneously for all GCSEs; it is probably better to restrict any such effort to within a subject.

There seem to be two opposing strategies advocated: either to use, in each question, the command phrase that most precisely states what is expected of a pupil, in which case several hundred phrases will be current in any one exam session, often with very slightly different nuances, or to use within each subject a quite limited set of phrases and try to 'teach' teachers and their pupils how to interpret each one in a particular context.

We are inclined towards the latter option, since we see the meaning of a command phrase as being a part of the student's schema, or expectation, of what a particular exam is likely to require, and so as something in which the teacher has a duty to prepare them. We think it unlikely that expert examiners will understand the subtleties of language in quite the same way as 15 year olds, or even that all 15 year olds will share the same understanding of them. We are sure, however, that the issue should be explored empirically: **what matters is not what command words 'mean' but what pupils *do* in response to them**. For example, it may be that, as with much younger children, the actual command word used will have little effect on pupils' behaviour and they will respond more to the content and how they have met it in the past. We recommend some **empirical investigation of the effect of using different command words** in the same question.

In the spirit of our 'grounded theory' methodology we offer the following suggestion of a three-step approach to systematising the use of command words:

1 **Examining teams should compile a list of command words** that satisfy their needs – we think that 20-30 words or phrases might suffice in each case. Lest this become too restrictive, an initial list should be treated only as a draft. This list can then be glossed, preferably with illustrative examples of questions and mark schemes rather than formal definitions, with a view to making it part of the assessment specification.

2 **Teams working in related areas**, or in different awarding bodies in the same subject, **should share these glossaries** at an early stage, with the implication that any egregious differences should be resolved. This will help examiners to appreciate the problems that they present to the child, who has only one mind with which to approach these many disciplines.

3 Ultimately, **a national project should collate the lists into a national encyclopaedia,** which will show how each word on the lists is being used in different disciplines. There should be no compulsion on any examining team to conform to national norms if they can justify their particular uses of a word in terms of the academic discourse of specialists in their domain, but the encyclopaedia will inevitably apply a mild pressure towards consistency.

An important element of any glossary should be a linking of each command word to a specific level in a cognitive hierarchy. Rather than Bloom's taxonomy we suggest that a fairly simple hierarchical scheme, (such as that by Peel and Sutherland that we discussed in Section 9.1.7) – which was developed to describe the range of cognitive skills demonstrated in British examinations – should be used to ensure that everyone – question writers, candidates, examiners and teachers – share a common understanding of the significance of each command word used.

To develop our comments on 'explain', the commonest of all command words, we also recommend, **a quantitative study of how 'explain' is used, and how the Reponses to it are evaluated, across all subjects**.

10.7 Demands

The relationship between command words and the cognitive demands of a question would seem to be straightforward, but it is not. We have seen cases, for example, where full marks can be gained in an 'explain' question by mentioning points that do not go beyond description. In such a case it is hard to decide what the actual level of cognitive demand is – should it be what the command word asked for, what the pupils actually tried to do, or what the mark scheme required for full or partial marks. When there is concern that the level of demands in examination is not being maintained it would be wise to consider the effect of command words and mark schemes on demands more carefully.

In Section 9.1.7 we mentioned the command phrase ‘give *n* reasons’, which seems to be a less demanding form of ‘explain’. Our data suggest that this form is becoming increasingly popular – it was used in 1.8% of our question sample in 2002, in 2.8% in 2005 and in 6.3% of the 2006 questions. Given that a superficial analysis would probably classify these as being just as demanding as ‘explain’ questions, we suggest **a quantitative study to consider what impact changes like this have on overall levels of demand**.

Where the overall level of demand is to be judged in terms of the command words or cognitive processes required to answer the questions, then we recommend that **serious consideration be given to using a scheme based on the work of Peel and Sutherland on British examinations,** rather than on Bloom’s American system.

10.8 Scripts

Our approach to this project was based on the psychology of exam questions. Over the years we have built up a model that predicts how students will respond to the questions they are posed, using our experience of how particular features affect their thinking. But these are still predictions, and not as reliable as direct evidence. In a study like this, the conclusions rely to some extent on the accuracy of the predictions. To increase the impact of the study **we recommend empirical studies of scripts to confirm where we predict problems, and to show how improved mark schemes or questions would have functioned better**.

10.9 IT implications

A shift to on-screen questions does not *necessarily* change any of our conclusions. Students will still read a question, probably in words with the aid of pictures, and will still select or ‘write’ their response, probably with a pen or keyboard. Our model of the question answering process will still apply, and help examiners predict the range of responses they will receive. There are some likely changes in practice, however, that will affect our recommendations: most of them are more likely to help than to handicap us.

We have already noted the value of **question level data** as feedback in the professional development of examiners. Awarding bodies should look to exploit this new resource as soon as possible.

One consequence of on-screen marking is likely to be a reduction in the number of questions that each marker will be asked to deal with; since it is no longer necessary to assign whole papers to each marker it makes sense to send each one only questions that they are particularly skilled at assessing. In these circumstances, it will be worthwhile to use **more elaborate mark schemes** that go further than at present in explaining to the marker what the examiner hoped to achieve with the question. A greater investment in fewer questions is entirely in the spirit of OSCA theory.

OSCA focuses on the mental behaviours that we expect students to demonstrate, rather than on the body of knowledge we expect then to master. Judicious use of IT in examining will expand the range of assessable behaviours, bringing them closer to the range that teachers want to develop in pupils. To achieve this it should be used where – and only where – it allows pupils to carry out more authentic mental processes. In Geography this might mean giving pupils access to larger and more varied maps or tables of data, or in Business Studies they might be required to evaluate more complex sets of accounts or of marketing data. The use of IT in more authentic Design and Technology tasks is obvious. In any such case, the assessment aim should still be to manipulate the students’ minds so that they are indeed doing the things we want them to show us they can do: if the task is more authentic then this will be easier to achieve.

10.10 The quality of mark schemes

We have already noted that the quality of the questions in examination papers has improved over recent decades. We now see more serious problems in mark schemes than in the questions. We are not aware of any prior scheme that outlines the features that make for quality in a mark scheme for very or semi-constrained questions, which probably reflects a general bias that needs attention.

We suggest that **particular emphasis be given by all awarding bodies to improving the quality of mark schemes**; we believe this would be the quickest and most effective way to achieve a rapid improvement in the quality of GCSE assessment.

A similar bias applies to research. We further suggest **that research be commissioned into both the procedural and psychological aspects of how markers make their judgements**.

REFERENCES

Achinstein, P. (1983) *The nature of explanation.* Oxford University Press.

Ahmed, A, Pollitt, A & Rose, L (1999). *Should oral assessment be used more often?* BERA, Leeds. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/publication.2004-09-15.6164501187

Ahmed, A & Pollitt, A (2001). *Science or Reading?: how students think when answering TIMSS questions*. IAEA, Rio de Janeiro. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/IAEA2001APAA

Ahmed, A & Pollitt, A (2007). Improving the quality of contextualised questions: an experimental investigation of focus. *Assessment in Education: principles, policy & practice*. 14, 201-233.

Anderson, J. R. (1983) T*he Architecture of Cognition*. Lawrence Erlbaum Associates, New Jersey.

Biggs, J. B. and Collis, K. F. (1982) *Evaluating the quality of learning: The SOLO taxonomy*. New York and London, Academic Press.

Bloom, B. S., Engelhardt, M. D., Furst, E. J., Hill, W. H. and Krathwohl, D. (1956) *Taxonomy of educational objectives: The cognitive domain*. New York, McKay.

Brown, A. L., & Day, J. D. (1983) Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 1983, 22(1), 1-14

Chandrasekeran, B. (1988) Explanation: the role of control strategies and deep models. In J. Hendler (Ed.), *Expert Systems: The user interface.* Ablex Publishing.

Charmaz, K. (2006) *Constructing grounded theory. A practical guide through qualitative analysis.* Sage, London.

Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J. & Cohen, G.M. (1997) Changes in Memory Awareness During Learning: The Acquisition of Knowledge by Psychology Undergraduates. *Journal of Experimental Psychology: General. 126, (4), 393-413.*

Crisp, V., Sweiry, E., Ahmed, A. and Pollitt, A. (in press). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research*.

Dey, I. (1999) *Grounding grounded theory. Guidelines for qualitative inquiry*. Academic Press, London.

Evans, J St B.T. (1989) *Bias in Human Reasoning: Causes and Consequences*. Hove: Lawrence Erlbaum.

Glaser, B.G. And Strauss, A.L. (1967) *The discovery of grounded theory: Strategies for qualitative research*. Aldine, New York.

Haladyna, T. M., Downing, S. M. and Rodriguez, M. C. (2002) A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement In Education*, 15, 309-334. See attached file.

Hempel, C. (1965) *Aspects of scientific explanation.* Free Press.

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

McLean, L.D. (2005) Organizational culture's influence on creativity and innovation: a review of the literature and implications for human resource development. *Advances in Developing Human Resources*, 7, 226 - 246.

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

Messick, S. (1990) *Validity of Test Interpretation and Use.* Research Report 90-11. Princeton, NJ: Educational Testing Service.

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, 3, 395 – 422

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

Piantanida, M., Tananis, C.A. and Grubs, R.E. (2004) Generating grounded theory of/for educational practice: the journey of three epistemorphs. *International Journal of Qualitative Studies*, 17, 3, May-June, 325-346.

Pollitt, A, (1991) Giving students a sporting chance: assessment by counting and by judging (reprint). In Alderson, JC & North, B (eds) *Language Testing in the 1990s: The Communicative Legacy*, Macmillan.

Pollitt, A, Entwistle, NJ, Hutchinson, CJ & de Luca, C (1985) *What makes Exam Questions Difficult?* Edinburgh, Scottish Academic Press.

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H. and Bramley, T. (1998) *The effect of structure on the demands in GCSE and A Level questions*. Report to QCA.

Pollitt, A & Ahmed, A (1999) *A new model of the question answering process*. IAEA, Bled, Slovenia. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/IAEA1999APAA

Pollitt, A & Ahmed, A (2000) *Comprehension Failures in Educational Assessment*. ECER, Edinburgh. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/ECER2000APAA

Pollitt, A & Ahmed, A (2001). *Understanding students' minds: how to write more valid questions*. AEA-Europe, Krakow, Poland. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/AEAEurope2001APAA

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

Roid, G. H., & Haladyna, T. M. (1982). *Toward a technology of test-item writing.* New York: Academic.

Salmon, W. (1970) Statistical explanation. In Colodny, R. (Ed.) *The nature and function of scientific theories*. University Pittsburgh Press.

Shalley, C.E. and Gilson, L.L. (2004) What leaders need to know: a review of social and contextual factors that can foster or hinder creativity. *The Leadership Quarterly*, 15, 33-53.

van Fraassen, B. (1980) *The scientific image.* Clarendon Press, Oxford.

Weiner, J.L. (1989) The effect of user models on the production of explanations. Ellis, C. (Ed.) *Expert knowledge and explanation: the knowledge-language interface.* Ellis Horwood.

Appendix 1

Command words and phrases used in British certificate examinations

Auxiliary verbs

do/does?
is/are/was?

Interrogatives

how?
what?
when?
where?
which?
why?

Verbs

account for
act
adapt
administer
advise
agree
allocate
analyse
anticipate
apply
appraise
arrange
articulate
assess
associate
balance
break down
calculate
change
chart
circle
cite
classify
collaborate
collect
combine
comment on
compare
compile
complete
compose
compute
conclude
connect
consider
construct
contrast
convert
convince
copy
correlate
cost
create
criticise
decide
defend
define
demonstrate
derive
describe
design
determine
develop
devise
diagram
differentiate
discover
discriminate
discuss
distinguish
divide
draw
enumerate
establish
estimate
evaluate
examine
experiment
explain
express
extend
facilitate
find
focus
formulate
generalise
give
grade
group
identify
if . . . would
illustrate
indicate
infer
integrate
interpret
intervene
invent
join
judge
justify
label
list
make a case
make a study
match
measure
modify
name
negotiate
normalise
obtain
order
outline
paraphrase
paste
persuade
plan
point out
predict
prepare
prioritise
produce
prove
quote
rank
read
rearrange
record
reduce
refer
reframe
reinforce
relate
reorganise
report
reproduce
restate
retell
rewrite
select
separate
show
sketch
sketch
solve
specify
speculate
state
structure
subdivide
substitute
suggest
summarise
support
tabulate
teach
tell
test
trace
transfer
use
validate
verify
write

Phrasal verbs, etc

give details
give evidence
give example
give reason
how accurate?
how does?
how effective(ly)?
how far?
how important?
how many?
how much?
how significant?
how successfully?
how valid etc?
to what extent?
what evidence is there?
what do you understand by?
write down
write algorithm (comp)
write essay

Multiple verbs

compare & contrast
describe & give reason(s)
describe & suggest
select & give reason(s)

Complete sentence

< incomplete sentence>

Appendix 2

Command words and phrases used in the Business Studies papers in this study

Command word	N	%	Command word	N	%
explain	176	29.7	comment on	3	0.5
give reason	41	6.9	consider	3	0.5
describe	34	5.7	define	3	0.5
calculate	31	5.2	give details	3	0.5
discuss	30	5.1	give example	3	0.5
what?	29	4.9	identify	3	0.5
<complete sent	23	3.9	write (down)	3	0.5
state	20	3.4	how (other)?	2	0.3
evaluate	19	3.2	how many?	2	0.3
suggest	16	2.7	if . . . would	2	0.3
which?	16	2.7	indicate	2	0.3
why?	13	2.2	report	2	0.3
what meant?	12	2.0	use	2	0.3
advise	11	1.9	allocate	1	0.2
select	11	1.9	compare	1	0.2
give	10	1.7	decide	1	0.2
how does?	9	1.5	distinguish	1	0.2
outline	9	1.5	how much?	1	0.2
name	8	1.3	how successfully?	1	0.2
complete	7	1.2	is/are?	1	0.2
justify	6	1.0	label	1	0.2
list	6	1.0	match	1	0.2
assess	4	0.7	order	1	0.2
what understand?	4	0.7	select & give reason	1	0.2
analyse	3	0.5	when?	1	0.2
				593	100.0