THE EFFECTS OF STRUCTURE ON THE DEMANDS IN GCSE AND A LEVEL QUESTIONS

Final Research Report

U.C.L.E.S.

Alastair Pollitt
Sarah Hughes Ayesha Ahmed
Hannah Fisher-Hoch Tom Bramley

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions*. Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.

The Effects of Structure on the Demands in GCSE and A Level Questions

CONTENTS

PREFACE 2

1. Chapter 1 BACKGROUND 4

1.1 The Research Question 4
1.2 Related Literature 5

2. Chapter 2 METHODOLOGY 12

3. Chapter 3 GEOGRAPHY 16

3.1 Geography GCSE 16
3.2 Geography A Level 34
3.4 Summary 47

4. Chapter 4 CHEMISTRY 49

4.1 Chemistry GCSE 49
4.2 Chemistry A Level 58
4.3 Summary 64

5. Chapter 5 HISTORY 66

5.1 History GCSE 66
5.2 History A Level 79
5.3 Summary 91

6. Chapter 6 SUMMARY OF SUBJECT FINDINGS 92

7. Chapter 7 THE EFFECT OF QUESTION STRUCTURE ON RELIABILITY 97

8. Chapter 8 STRUCTURE IN MATHEMATICS A LEVEL 101

8.1 Structure 101
8.2 Question Difficulty 105
8.3 Demand 108
8.4 Maths A Level Conclusion 110

9. Chapter 9 CONCLUSIONS 111
GLOSSARY OF TERMS 127
REFERENCES 128

APPENDICES
APPENDIX 1 Geography Questions and Mark Schemes
APPENDIX 2 Chemistry Questions, Mark Schemes and Data
APPENDIX 3 History Questions and Mark Schems
APPENDIX 4 Draft Guidance Materials

THE EFFECTS OF STRUCTURE ON THE DEMANDS IN GCSE AND A LEVEL QUESTIONS

PREFACE

The aim of this project is to investigate the effects of the structure of exam questions on the demands made on students. A structured question is either a large task broken down into parts or a series of questions written on a common theme. **Structure** has been used in exam questions in the past with the aim of increasing construct validity by reducing non-subject demands such as linguistic skills, and to increase reliability, as more specific mark schemes can be written for structured questions. In some subjects structuring has been used to control difficulty, by writing structured questions for lower tier papers to give those candidates more support. Empirical investigations were carried out asking whether structured questions are less demanding than unstructured ones, and addressing the nature of **demands** and how these are perceived by examiners and students in questions of different structure.

Examiners wrote questions for GCSE and A Level Geography, Chemistry and History at three levels of structure, whilst keeping the content as similar as possible across versions. The report describes the effects of structure on (i) the difficulty of questions (or the performance of students), (ii) the demands perceived by examiners in the questions, (iii) examiners' perceptions of the quality of the responses, and (iv) the demands perceived by students in the questions.

For difficulty, average marks gained by students on the three differently structured versions of the same question were compared. Examiners were also asked to rate each question on four scales of perceived demands labelled Complexity, Abstraction, Resources and Strategy (based loosely on Edwards and Dall'Alba, 1981). Examiners also gave ratings on the same four scales to the overall quality of the responses to the differently structured questions. To extend this further, examiners were interviewed using the construct elicitation part of the Repertory Grid Technique (Kelly, 1955) in order to understand how they view demands in their subject.

The concluding data consisted of interviews with students, in which they were asked to compare three differently structured versions of the same questions and talk about which were more demanding and why.

It was found that in Geography, structured questions were easier than unstructured questions in most cases, though there were questions in which examiners introduced unintended demands into the more structured versions. It must be remembered here that "easier" simply means that students scored more marks; in an examination context we would expect the grade boundaries to be raised to compensate for this. In Chemistry, structured questions were easier than unstructured questions and made fewer demands almost all of the time. In History, however, students performed equally at all levels of structure, and performance was principally influenced by students' expectations of the sort of questions they would be asked. The A Level students in particular, preferred unstructured questions as they had been prepared for these.

Two small studies are also included in this report. Chapter 8 describes differences in structure and demands between 1986 and 1996 A Level Mathematics papers, and Chapter 7 looks at how structure affects the marking and awarding process.

Five types of structuring were used: breaking down the question into sub-parts; providing prompts to the content of an answer; providing a strategy for answering; changing the response format; and changing the wording of the question. All, except the first of these, had an effect on performance and on demand ratings in terms of complexity and strategy. It can be concluded that structured questions are not always easier or less demanding than unstructured questions. The expectations and ability of students are important factors, as is the type of structuring and the nature of the skills assessed. Implications of these findings for the validity and reliability of assessment are discussed.

The report also includes preliminary ideas for guidance materials based on the findings of this project (see Appendix 4). These will provide help to question setters on the use of structure. Chapters 3, 4 and 5 of the report are subject specific chapters, but for non-subject specialists Chapter 6 contains a summary of the empirical findings from all three subjects.

1. BACKGROUND

1.1 The research question

The aim of the project was to investigate the effects of question style on 'demand' in a range of GCSE and A Level questions, where 'style' refers particularly to the kinds of changes that have been introduced in recent decades in what is seen as a move to 'more structured' questions. In an attempt to distinguish the various ways in which changes affect cognitive processing it is necessary to analyse the nature of 'structure' and the nature of 'demand' (here expanded to 'demands' throughout).

Extending Pollitt et al (1985), recent work in UCLES (e.g., Hughes and Fisher-Hoch, 1997) have identified a number of *sources of difficulty* (SODs) in GCSE exam questions. Some 'SODs' related in various ways to the structure of questions, notably those called 'Sequencing' (the sequence and interaction of the sub-parts of a question), 'Paper layout' (the physical organisation of the question and space for response) and 'Number of steps' (it was found that a large number of steps overload working memory and information was likely to be lost). Questions showing these features were rewritten to alter the nature of the SODs and the performance of experimental groups of pupils on the various versions was compared. Results showed that 'structure' clearly can affect the performance of candidates.

In this investigation, the work is extended to A Level, and the experimental manipulation focuses on the kinds of changes that constitute structure. Questions were manipulated from GCSE and A Level History, Chemistry and Geography. We considered the effects of these versions on: (i) question difficulty, (ii) the demands made on students (as intended by the examiners), (iii) the demands made on students (as observed by the examiners in their scrutiny of scripts) and (iv) students' perceptions of exam questions and the demands they make.

1.1.1 The use of structure in school examinations

Three reasons may be described for the introduction of structure to school examinations. First, it was intended to increase the validity of exams. Traditionally school exams used essay questions with a large problem space (Newell and Simon 1972); there were potentially many ways to answer the prompt, several of which might gain equal credit, and it was part of the pupil's task to decide how to tackle it. Furthermore, mastery of the subject could only be demonstrated through extended writing, posing severe problems for those candidates with poor composition or writing skills. The structuring of questions was increased to reduce the demands that were considered not to be part of the particular subject being examined and so to increase construct validity. For example, in science (MEG 1992) it was noted that although more open ended questions may be used, it was thought to be a risky assumption that the most able candidates in science are all good at English and communication.

Secondly, the use of structured questions could increase the reliability of marking. For essay examinations the variability of marking between different examiners is a major source of unreliability. Murphy (1982) stated that:

> '...the reason why more closely defined questions might be more reliably marked than free-response essay type questions are understandable in terms of the nature of the instructions which can be given to examiners in these different situations.' (Page 199).'

A structured question makes more specific demands, so that the mark scheme will be more specific, reducing inter-marker unreliability.

Thirdly, structure is used in examinations to control the difficulty of questions. (See, for example, papers on setting effective questions (SCAA 1992). The use of structure as support for candidates at lower grades is widely applied particularly in the GCSE where tiered papers aim to test the same or similar content across different ability levels. In science, for example, it was suggested that questions designed for level 10 (A*) '...should give candidates the opportunity to write about a problem in depth.' (SCAA 1992, page 17). This could be done by asking open ended questions in which candidates had to use extended prose. On the other hand 'questions which are finely structured with prompts telling candidates exactly what to do at every step' were likely to be inappropriate at the higher level. Thus adding structure would make questions more *accessible* while removing it would make them more *demanding*.

Structuring is used widely in GCSE and A Level examinations for the three reasons described above. Yet there is little research showing what effect structuring has on how candidates tackle questions and what demands structured and unstructured questions make on them. It is hoped that the results of this project will be of immediate value to examination question writers, and we intend that one significant outcome should be guidance materials for the writing process.

1.2 Related Literature

1.2.1 The nature of structure in school examinations

Wilmut (1979) described the typical features of structured questions: they are set within a defined skill or content area; they contain several question parts; those parts are linked together; and each question is asked in such a way as to specify the type of response required (so the mark scheme is sufficiently specific to reflect the structure of the question).

Structured questions have a number of benefits over open questions (Murphy 1982):

- There is better communication between examiner and candidate about what exactly is required
- Candidates can't rehearse whole answers
- Mark schemes can be more specific, thus increasing the reliability of marking
- The mark allocation for sub-parts helps candidates allocate their time

- Complexity can be increased incrementally throughout a question

The problem space of an unstructured question may involve a number of sub-goals that a candidate must identify and work through in order to reach the final goal of a complete answer. The sub-goals of a structured question are made explicit, and candidates are required to attempt, show workings for, and give an answer to each of the sub-goals as if they were separate questions.

Structured questions thus reduce the need for candidates to plan and monitor their work, as some of the strategy for answering a question is given by the steps of the question parts. Yet unstructured questions may allow pupils to avoid showing their ignorance, and can favour those who write with facility. Structured questions address the syllabus more specifically, allowing examiners to test more exactly the knowledge and understanding they think pupils ought to have. The dilemma for examiners is clear: structured questions can increase the examination's focus on skills and knowledge relevant to the particular subject and reduce the importance of more general skills, yet in doing this they may ignore some high level cognitive skills that everyone agrees should be valued. If the examination doesn't test them experience suggests that teachers won't teach them.

1.2.2 **Demands in school examinations**

Demands can be seen as requests that examiners make of candidates, to perform certain tasks within a question. There can be many or few demands in a question, and these demands may be complex or simple. Skills, knowledge, understanding and the ability to apply these are generally seen by examiners as the demands of exam questions. Given that there are many different sorts of demands in examination questions we choose in this report to talk mostly of *demands* rather than *demand.*

Sometimes demands are described euphemistically as 'opportunities', where 'extra' credit will be given to those who can meet them. Sometimes they are explicit in the question or task, but sometimes they reside in the mark scheme and it is the job of the teacher to ensure that pupils are fully aware of them. In recent years the use of grade and level descriptors in the more literary subjects has increased the awareness amongst pupils and teachers of the implicit demands residing in the mark scheme, and it has become clear that these demands in fact constitute a significant part of the syllabus.

Work in maths, however, (SRAC 1990) found that while skilled judges (for example examiners) recognise demands and tend to agree with each other about the general level of demands in questions, they could not identify the constituent parts of 'demand' nor could they explain <u>why</u> they found something demanding. It seems that although there was a shared understanding and conception of demand, a clear, explicit definition was lacking.

It is possible that this inability to be explicit about demands in exam questions might be less a fault of the judges than a fault with the mechanisms used for eliciting the judgements of demands from the judges. Two techniques from the psychological literature might profitably be used to identify demands in questions: Kelly's repertory grid technique and Edwards' scale of cognitive demands.

1.2.2.1 The Repertory Grid Technique

The repertory grid interview is a tool to help interviewees communicate their views and ideas using their own, meaningful language. It was developed by Kelly (1955) within the framework of his *personal construct theory*, a theory of personality development that is widely used in research and therapy today. The primary focus of personal construct psychology was upon the way individuals perceive their environment and the way they interpret what they experience. A developing child (and indeed an adult) is described as a 'scientist' constantly forming hypotheses in an attempt to understand and gain control of the world. In the best traditions of the philosophy of science the person employs as few and as simple hypotheses as possible. Thus we can expect a person in any given context to 'make sense' of what they see by sorting and classifying all of the phenomena experienced according to a few simple criteria. These are the *constructs* of the theory, since they do not have any independent existence in the world but are built up by the person from their experience. While we may be predisposed by evolution to build certain kinds of construct rather than others, Kelly emphasises the *personal* nature of the experience that goes into building them, and it is therefore clear that each person's repertoire of constructs will be different. The repertory grid technique is a method for making those constructs explicit. By controlled presentation of stimuli the person is provoked into revealing the criteria that are most salient for them at that moment, and from this the researcher (or therapist) can infer how the person perceives their experiences, even though the person may not be able to make the constructs explicit.

The method has been applied to education in several ways. Kremer-Hayson (1991) used it to elicit from teacher-managers their perceptions of good professional practice. Fisher et al (1991) applied the technique to course evaluation. Parsons et al (1983) showed how the grid technique could be used to make explicit the implicit models that teachers have of how children learn. Previous research has shown that examiner's conceptions of demands tend to be implicit, and the power of repertory grid interviews is that they can make the implicit explicit. This technique could therefore enable examiners to share their tacit knowledge.

1.2.2.2 The Scale of Cognitive Demands

Edwards and Dall'Alba (1981) developed and implemented a 'Scale of Cognitive Demand'. The scale was developed as an instrument for analysing secondary science lessons, materials and programs in Australia. It is intended to quantify the demands placed on the cognitive abilities of students. The conceptualisation of demands was derived from a range of learning

and thinking theories, including Bloom (1956), Bruner et al (1966), de Bono (1976), Gagne (1970), Taba (1962, 1967), Ausubel et al (1978) and the work of Piaget as interpreted by Novak (1977). The scale considered demands to have four dimensions: Complexity, Openness, Implicitness and Level of Abstraction. Six levels of demands were defined within each dimension, by a list of phrases and command words that were typically used in science textbooks and examinations, or that could be used to describe the processes students were required to carry out. For this study the scales were modified (see Hughes et al (1998) for more detail on the development of the scale) to enable the scale to be applied to subjects other than science. We found it necessary also to change from six defined levels to a 1-5 continuum with only levels 2 and 4 described verbally. The four dimensions are shown in the table below:

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

1.2.3 **The relationship between structure and demands**

Newell and Simon's (1972) work on the psychological concept of *problem space*, and Marton's (e.g., Marton & Saljo, 1976) treatment of *outcome space* provide useful methods for considering the relationship between the structure of exam questions and the nature and level of demands in them. For any given problem there may be many alternative paths from an initial state (the candidate reading the question) to a goal state (the candidate having written the perfect answer): the total set of these paths is called the problem space. The candidate clearly must 'have' all of the knowledge represented by at least one of these paths in order to solve the problem, and the skills necessary to traverse it; in practice a candidate is likely to have much more knowledge than that and will be faced with the need to select an appropriate subset of knowledge that will constitute an acceptable path. Examiners can help or hinder this process by providing the metaphorical equivalent of signposts that suggest routes to take or debar the candidate from following certain paths.

In our context problem space can be restricted in several ways. First, examiners may include visual clues showing what the question requires (for example through the use of bold type to direct candidates, or clear designation of the space in which an answer should be written). Second, the amount and type of information provided to candidates can restrict or enlarge the problem space (e.g., by giving a formula or diagram). Third, the problem space can be adjusted by giving clues to the best strategy for answering the question (for example, giving a line reference for a piece of text or telling the candidate the topic of the question). Explicit phrases such as 'an example you have studied' may greatly enlarge the problem space. Each of these features has been identified in our research on question difficulty and examiners have confirmed that they deliberately use such features as a tool to vary demands.

We can see how the outcome space of a question can be restricted or enlarged through the use of structure; a more structured question will usually reduce the problem space and a more open question will usually enlarge it. This suggests that structured questions will usually be less demanding than open ended questions. This explanation is in line with the current practice of examiners who use structure as a tool for varying the demands in exam questions, for example in maths it has been concluded that

> 'Open-ended questions often require verbal, rather than non-verbal responses, or are less structured rather than more structured. Either way, they are found to be more difficult by the less able candidates'. (SCAA 1992, page 21)

It would be over simplistic, however, to see difficulty as a simple function of the size of the problem space. Marton, like Kelly, emphasises the individual nature of understanding. Each pupil develops their own understanding of the subject as they study it, and in an examination they each bring this individual understanding to bear on the problems they are set. It is likely then that they will produce responses that differ qualitatively - different kinds of answer - as they will certainly approach many problems in their own particular way. This individual variation will interact with the definition of the problem space, so that it will actually be a

different size for each person, containing only those paths and potential paths that they perceive as relevant. We should perhaps consider the problem space from the candidate's point of view, as the set of all paths that each person believes may lead from initial state to goal state. Then we can ask how the various features of 'structure' might contribute to the definition of each candidate's personal problem space.

Through this review of the literature and our previous work in the area of question difficulty we can suggest a number of ways in which structured questions appear to be usually less demanding than unstructured or open ended questions:

- The need for extended writing is reduced. This reduces the (non-subject) demand that candidates articulate their knowledge in continuous writing.
- There is better communication between examiner and candidate about what exactly is required. For example, by asking more specific questions the skill or content area is more defined.
- The need to plan and monitor one's working and answer is reduced when answering a structured question.
- A structured question is likely to have a smaller problem-space.
- The mark allocation for sub-parts helps candidates allocate their time.

This research is principally concerned with investigating the accepted wisdom that structured questions make fewer demands than open questions.

1.2.4 **Structure and its effects on marking and awarding**

A further concern of this investigation is the effect of structured and unstructured questions on examiner's judgements of scripts. From Marton's discussion we should expect that a structured question will have a more tightly defined problem space, with less variation between candidates in its definition and in their outcome space.

Murphy (1977) proposed that there were fewer problems of inconsistency in marking in maths than other subjects. In further research it was found that maths is significantly more reliably marked than English (Murphy 1982). The differences between subjects could be due to a number of factors. One explanation could lie in the difference in structuring of English and Maths questions in 1982, with English questions being typically less structured than maths questions. However, the effect of structure was not one of Murphy's foci so we would be mistaken to attribute differences in marking reliability to structure.

Newton (1996) also found that Maths was more reliably marked than English. One explanation for this was that the degree of specificity of the question and mark scheme: 'When marking schemes are broken down to clarify precisely what each mark is being awarded for, we an expect the reliability of marking to be high. The highly detailed marking schemes for mathematics help explain the high degree of reliability obtained.' (p418).

Good and Cresswell (1988) found that a higher proportion of candidates reached the grade boundary on easy components (lower tiers) than on harder ones. They suggested that awarders may be applying standards to individual tiers which they had previously applied to a complete examination which assessed the whole ability range. It is also possible, though, that the differences in question type across tiers could have impacted upon examiners' awarding judgements. Lower tier papers tend to be more structured than higher tier ones. If extra structure makes papers more 'accessible' while less structure makes them more 'demanding' then we would expect borderline candidates to produce better performances on the lower tier. Then since judges in award meetings try to equate on the basis of *performance* rather than *ability* it is entirely predictable that borderline candidates should do better on lower tiers.

1.3 **Background summary**

Structure is widely used in school examinations to increase validity by reducing the non-subject demands made on candidates and to control the demands questions make on candidates. However, little research has addressed the assumption that structuring questions has the desired effects. This study aims to establish the effect of structuring questions on examining. Four areas were of particular interest:

- The effect of structure on candidates' performance.
- The demands that examiners consider structured and open questions to make on candidates.
- Candidates' reactions to and perceptions of differently structured questions
- The impact of structure on examiners' judgements of scripts.

2. METHODOLOGY

The main part of the study concerns examining in three subjects - History, Geography and Chemistry - and at two levels - GCSE and A Level. A variety of methods were used to explore the consequences of changing the degree of structuring in questions, and they are outlined in this chapter. In general the same procedures were used in each of the six sub-studies; where there are any deviations from this pattern they will be mentioned in the relevant subject chapter.

All of the procedures depended on the **questions** written for the study, and the first section describes how these were generated. An obvious first step then was to see if the various versions differed in **empirical difficulty**. Since the study is concerned with the **demands** that candidates face, a method was developed for collecting examiners' impressions of these demands in the various question versions. Since this procedure used rating scales from another context (Australia) a further procedure was designed to elicit the examiners' own **constructs** for rating questions. Two observational elements were added to these procedures. In the first we tried to understand the effect of structure on the process that occurs as **examiners judge** the value of responses, taking into account the nature of the particular question asked. Finally, we recorded and analysed **interview protocols** as pupils considered how to answer the questions. Each of these procedures is described below, in general terms that apply to each of the six sub-studies.

2.1 Question writing

Six examiners (one for each of GCSE and A Level in each subject) were asked to write questions at three levels of structure: Structured (ST), Semi-structured (SS) and Unstructured (US). The question writing took place during a two day residential meeting, after examiners were fully briefed about the background and aims of the project. In each case the questions were designed to conform as closely as possible to existing UCLES examinations. The various questions were put together to make two papers, each lasting 50 minutes, for trialling in schools with the minimum of disruption to the normal timetable. There were three versions of each paper, each version containing a mixture of ST, SS and US questions. This mixed format was chosen in preference to complete papers of US, SS and ST questions partly to ensure that each version would take the same length of time to complete, and partly to ensure equal 'morale' if a particular type of question should turn out to be unfamiliar or off-putting.

The examiners were instructed when writing questions at different levels of structure to try to make the content of the questions as similar as possible, so that we would be able to interpret any differences in performance among the versions as being due to the differences in structure. This should also mean that examiners would be able to mark all of the versions using the same mark scheme. The extent to which it was possible to change the structure under this constraint varied among the three subjects, and details are given in later chapters.

2.2 Empirical estimation of difficulty

In an initial pilot phase 10 candidates took both papers for each subject at GCSE and A Level, to make sure that the questions and mark schemes were working as intended. A few changes were made in the light of this pilot, and then the main studies were carried out. Schools using the relevant UCLES syllabuses were contacted and invited to take part, until enough had agreed to ensure that a reasonable number of candidates (approximately fifty) would be taking each version of each question. In A Level History this was a difficult task, since in the examination candidates can choose to answer any 3 out of 25 essay questions. They are likely to be prepared for up to 6 questions, but which questions in which domains of History will depend on the school. The final numbers of schools and pupils taking part in the trial are shown below:

Subject	Number of schools	Number of pupils		version
		Paper 1	Paper 2	
GCSE Chemistry	2	44	43	A
		44	45	B
		44	44	C
A Level Chemistry	6	64	55	A
		63	56	B
		59	59	C
GCSE Geography	2	39	50	A
		45	55	B
		42	57	C
A Level Geography	2	55	60	A
		62	63	B
		67	68	C
GCSE History	2	79	70	A
		57	55	B
		57	54	C
A Level History	37	110	116	A
		119	112	B
		108	121	C

Testing took place in January 1998, and all of the pupils were intending to take the corresponding exam in June. An estimated grade for each pupil was also collected, primarily to enable us to check that we had an approximately equal distribution of ability in each of the different versions within a subject. The papers were marked by the examiner who had set the

questions and the results were reported back to the schools. The statistical analyses reported in the next three chapters are of the raw question marks awarded by these examiners.

2.3 Ratings of demands

As described in the previous chapter, the Scale of Cognitive Demands was adapted for the purposes of this study. The examiners were asked to rate the questions on four dimensions of demands - Complexity, Resources, Abstraction and Strategy. In some cases the examiners adapted the scale slightly to meet the preferences of their particular subject; these modifications are described in the following chapters. After this, the examiners were also asked to rate their overall impression of the responses to the questions, using the same scales. One of our concerns has been to distinguish, where appropriate, between the demands that examiners planned into the questions and the demands that actually operated when pupils tried to respond. Any mismatch here will relate to the concept of test validity.

The information from these ratings of demands was used in several ways: to see how ST, SS and US questions were perceived by examiners to vary in their demands; to see how the dimensions of demands related to the levels of structure; and to see if variations in performance on different levels of structure were matched by variations in rated levels of demands.

2.4 Rating construct elicitation

The dimensions of demands in the Cognitive Demands scales were 'imposed' on the examiners to a certain extent, although they were allowed to modify them to suit their subject. We also employed a version of the initial phase of Repertory Grid Analysis as a more fundamental approach to understanding the process of judging the quality of responses. The intention here was to try to get at the examiners' own ways of construing 'demands' in their subject - in the language of Kelly's theory the 'constructs' which form their concept of attainment. Although each examiner might be expected to have their own unique way of construing the dimension of attainment it seems reasonable - even essential for our purposes - to expect substantial correspondence between examiners used to a particular syllabus and its demands.

There are several methods regularly used for construct elicitation; in our case examiners were repeatedly shown 'triads' of three questions, chosen from a small subset of the total number of questions, and asked to say how two were similar to each other and different from the third. Qualitative analysis extracted from these data a set of constructs that typified each examiner's approach to construing attainment. These were then formalised into a small set of dimensions, and the examiners were then asked to rate all of the questions using these, their own, constructs.

2.5 Examiners' Judgements

In the final procedure applied to the examiners, 'naive' examiners (i.e. experienced but unfamiliar with the aims or methods of this project) were shown sample responses *on the same mark* from each of the different versions of a question, and asked to give their impression of the level of performance on each. In general they re-marked the responses, without knowing what mark the responses had originally been given. There were two aims: the first was to try to observe (albeit with a very small sample) how examiners 'allow' for the nature of the question when forming judgements about scripts. Secondly, we hoped to get some idea of the 'reliability' of ST, SS and US questions, both in terms of how well the mark and judgement of the naive examiner agreed with the first examiner and through the comments of the naive examiners as they compared the various responses.

2.6 Interviews with pupils

Our last source of evidence involved the students themselves. In each subject a few pupils were interviewed using 'paired verbal reports'. The aims were to find out first how they spontaneously structured the US question, and second what their reactions were to ST, SS and US questions. In this method two candidates worked on the same question at the same time. After an initial few minutes of individual work, planning outline answers to a question, they were asked to confer with each other to agree on an outline answer using the best bits of their individual answers. Previous experience with research into pupils' work processes has shown that pupils are able to talk more freely to each other than to a researcher, and that we are able to infer from a transcript of their discussion how they understood the question and how they structured their answer. At times the interviewer asked them to confirm what they had said, and also asked more direct questions, concerning for example which version they found easier and why; which versions were helpful or unhelpful or which version they preferred.

2.7 Note

The next three chapters describe in detail the results of the empirical studies in each subject - Geography, Chemistry and History. These results are then drawn together in Chapter 6, and the reader who is not interested in the subject specific details may move directly to that chapter.

In these chapters the following abbreviations are often used:

US = 'Unstructured', SS = 'Semi-structured', ST = 'Structured'.

3. GEOGRAPHY

3.1 Geography GCSE

3.1.1 *Structuring the questions*

The questions used in the three versions of the Geography GCSE papers were based on the Summer 1996 MEG Geography Syllabus 3 (1588) examination. The examiner re-wrote the original GCSE questions at three different levels of structure using a number of different techniques (see Appendix 1 for questions and mark schemes). In some questions structure was imposed by breaking down the question into sub-parts. The aim of this was to structure the process of answering the questions. However the questions all included some structuring of content. This was achieved in a variety of ways, sometimes by changing wording in the question. Question 1a(i), for example, in the structured version read[§] :

> *Use the appropriate letter and number to identify the square containing Koblenz, the largest settlement on the image.*

the semi-structured question read:

> *Give the co-ordinates for the square containing Koblenz, the largest settlement on the image.*

and the unstructured version read:

> *Give an accurate location for Koblenz, the largest settlement on the image.*

Another way structure was manipulated was by changing command words such as describe and explain. For example, in Question 1(c), the structured question asked students to:

> *(i) Describe what Fig. 2 is predicting many people will do.*
> *(ii) Suggest how two of the following might explain your answer in (i):*
> *Employment opportunities*
> *Standards of living*
> *Environmental factors*

whereas the semi-structured question read

> *(i) Describe what Fig. 2 is predicting many people will do.*
> *(ii) Suggest two reasons such as economic and social factors, to explain your answer.*

and the unstructured question read:

> *Explain and give two reasons for what Fig. 2 was predicting would happen.*

Prompts such as those in the structured version of Question 1c(ii) above were used in the structured version of all four questions. Highlighting was also used, with key words typed in bold in the structured questions.

In Question 3(b) students were given a table containing grid references and altitude in the structured version, whereas they were told to use evidence from the separate map in the semi-

---

§ Question text is displayed in italic script here; it was not in italics in the question papers.

structured and unstructured versions. This variation in resources may have an effect on the question content and on the process of answering it.

The effects of the structuring of each question are discussed in the next section.

3.1.2 *The effects of structure on difficulty*

3.1.2.1 Distribution of students across versions

The distributions of students' forecast grades across the six versions did not vary significantly. The next table shows the numbers of students at each predicted grade who took each version. Students attempted four questions, with Questions 1 and 2 forming Paper 1 and Questions 3 and 4 forming Paper 2.

	Paper 1			Paper 2		
Predicted Grade	Version A	Version B	Version C	Version A	Version B	Version C
A	10	9	2	12	8	9
B	8	5	6	11	9	6
C	4	6	15	9	16	9
D	8	6	8	9	7	6
E	6	10	3	6	4	10
F	3	3	1	4	3	6
G	2	3	2	4	1	6
U	0	1	2	0	2	1
X	0	1	0	0	0	0
Total	41	44	39	55	50	53
Mean grade	5.8	5.2	5.3	5.7	5.7	5.0

In the following report the six groups will be considered equivalent, since the mean grades do not vary significantly. This meant we could be reasonably confident that differences in performance between differently structured versions of the same question were due to the differences in structure and not ability.

3.1.2.1.1 *Question 1*

The table below gives the mean marks of students on Question 1 for each version.

Question 1	Structured	Semi-structured	Unstructured
Mean mark	9.8	5.0	6.1
Maximum mark	20	20	20

Students performed significantly differently on the three versions on all of the sub-parts of Question 1. A detailed breakdown of performance on each part of this question and the effects of structuring can be seen in the following table.

Q 1	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
a(i)	ST & SS equal		US	$F_{(2,123)}$ = 6.67,p<0.005	letter & number (st) vs co-ordinates (ss) vs accurate location (us)	Everyday term used in technical sense
a(ii)	ST	US	SS	$F_{(2,123)}$ = 5.81,p<0.005	Where rivers join (st) vs confluence (ss) vs site (us)	Technical term
b	ST	SS	US	$F_{(2,123)}$ = 11.48,p<0.005	'similarities & differences'(st) vs 'different patterns of land use'(ss) vs 'patterns of land use' (us)	Specifying organisation of answer
c	ST	SS	US	$F_{(2,123)}$ = 13.94,p<0.005	Describe and suggest (st, ss) vs Explain (us)	Command words
d	ST	SS	US	$F_{(2,123)}$ = 6.05,p<0.005	Prompts (st) vs No prompts (ss, us)	Prompts specifying organisation and content of answer

3.1.2.1.2 *Question 2*

The next table gives the mean marks of students on Question 2 for each version.

Question 2	Structured	Semi-structured	Unstructured
Mean mark	6.5	6.8	5.7
Maximum mark	20	20	20

In all parts of this question, structure did not have a significant effect on difficulty. Some prompts were given in the structured version of the question, but these did not have much effect. The examiner thought that the structured version seemed daunting to students because there was more to read.

Q 2	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
	equal			$F_{(2,123)}$ = 0.83,n.s.	Prompts (st) vs No prompts, referring to graph (us)	None
	equal			$F_{(2,123)}$ = 0.08,n.s.	Explain (st) vs Suggest (ss,us)	None
	equal			$F_{(2,123)}$ = 1.43,n.s.	Two-part question (st) vs Single task (ss, us)	None
	equal			$F_{(2,123)}$ = 1.00,n.s.	Three parts, comment on success and/or failure (st) vs Comment on outcome (ss, us)	None

3.1.2.1.3 *Question 3*

This table gives the mean marks of students on Question 3 for each version.

Question 3	Structured	Semi-structured	Unstructured
Mean mark	5.8	5.6	5.7
Maximum mark	20	20	20

Again, in all parts of this question structure did not have a significant effect on difficulty.

Q 3	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
a	equal			$F_{(2,123)} = 2.61$,n.s.	Information given (st) vs Prompts (ss) vs No guidance (us)	None
b	equal			$F_{(2,123)} = 2.27$,n.s.	Data table + prompts (st) vs No guidance (ss,us)	None
c	equal			$F_{(2,123)} = 0.89$,n.s.	Two-part question + prompts (st) vs Single task, no prompts (ss, us)	None
d	equal			$F_{(2,123)} = 0.45$,n.s.	Three-part question + prompts (st) vs Two-part question + prompt (ss) Single task, no prompts (us)	None

3.1.2.1.4 *Question 4*

The next table gives the mean marks of students on Question 4 for each version.

Question 4	Structured	Semi-structured	Unstructured
Mean mark	7.8	5.3	6.1
Maximum mark	20	20	20

Parts (a), (c) and (d) of this question showed a significant effect of structuring.

Q 4	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
a	ST	US	SS	$F_{(2,123)} = 3.16, p<0.05$	Suggest what areas have in common (st) vs Describe type of area (ss) vs Describe pattern (us)	Generate theory rather than discover it via task
b	equal			$F_{(2,123)} = 2.02$,n.s.	Given kinds of area (st) vs What two areas (ss) vs Describe pattern (us)	None
c	ST	US	SS	$F_{(2,123)} = 3.58, p<0.05$	Prompts (st) vs One suggestion (ss) vs No suggestions (us)	Misleading wording in ss version
d	ST	US	SS	$F_{(2,123)} = 10.60, p<0.005$	Three parts, prompts (st) vs Two parts (ss) vs Single task (us)	Prompts specifying organisation of answer Breaking down into sub-parts

To summarise, the ways of structuring the questions that affected the difficulty of the different versions were:

1) The use of an everyday term in a technical sense
2) The use of a difficult technical term
3) Specifying the organisation of the answer required
4) Providing cues to the content of an answer, changing the requirement from production to recognition

5) The use of different command words such as describe and explain, changing the nature of the answer required

6) Breaking down task into sub-parts

7) Asking students to come up with a theory rather than leading them into discovering it via a specific task

8) Misleading wording of questions resulting from an attempt at structuring

This list illustrates the cognitive demands that were found in the differently structured questions. In the next section the examiner's perception of the level of demands in the questions and in the responses will be considered.

3.1.3 *Examiner's rating of demands*

3.1.3.1 Scale of demands

The Geography examiners worked together on their definitions of what each dimension in our adapted version of the Scale of Cognitive Demands (Edwards and Dall'Alba, 1981) meant in Geography. The scale that the examiners used is shown below.

	1	2	3	4	5
Type of Demand **Complexity** *The complexity of each component operations or ideas and the links between them*		**Types of questions showing lower demands** Description Small scale - local/personal Local/familiar Short sentences Human/ everyday	↔	**Types of questions showing higher demands** Analysis and synthesis Large scale - beyond personal experience Europe → LEDC Remote/ unfamiliar Long sentences Physical past and future	
Resources *The student's use of the data/information given*		Told where to look Simple resource Designed for question Topographic/isomorphic	↔	Not told where to look Complex resource Not designed for question Map transformation	
Level of Abstraction *The extent to which the student is required to deal with ideas rather than concrete objects or phenomena*		Everyday language Sentences Concrete - today Deals with today Idiography Locations	↔	Technical language mathematical formula imaginary - past/future projects forward and back pattern/distribution	
Strategy *The extent to which the student is required to devise and monitor a strategy for tackling the question and organise the information to be communicated in the answer*		Series of steps Deals with resources one at a time Identifies simple links between two ideas Makes decision with some evidence. elects case study with help/specification	↔	Essay with title and a sequence Selects from several versions Identifies several links between three or more ideas Makes decision, gives evidence, discusses and qualifies. Selects case study to criteria	

3.1.3.2 Examiner's ratings of demands

3.1.3.2.1 *Question 1*

The examiner rated each part of Question 1 on how demanding it was on each of the four dimensions of the scale. The ratings are shown below.

Statistically, students performed better on the structured question than on the other two versions for all sub-parts of this question. As the examiner has rated this question at the same level of demand in all three versions for Resources and Abstraction, we can conclude that the demands in this question were related to the complexity of the question and the strategy students had to use.

Demand	**Level of structure**	**a(i)**	**a(ii)**	**b**	**c**	**d**
Complexity	st	low	medium	medium	medium	medium
	ss	medium	medium	medium	medium	medium
	us	high	high	high	medium	medium
Resources	st	low	low	medium	medium	medium
	ss	low	low	medium	medium	medium
	us	low	low	medium	medium	medium
Abstraction	st	low	low	medium	medium	low
	ss	low	low	medium	medium	low
	us	low	low	medium	medium	low
Strategy	st	low	medium	low	low	medium
	ss	medium	medium	medium	medium	high
	us	high	medium	high	high	very high

3.1.3.2.2 *Question 2*

Each part of this question was rated as follows.

Demand	Level of structure	a	b	c	d
Complexity	st	medium	low	medium	medium
	ss	medium	medium	medium	high
	us	high	high	high	high
Resources	st	low	medium	medium	high
	ss	low	medium	medium	high
	us	low	high	medium	high
Abstraction	st	low	low	low	low
	ss	low	low	low	low
	us	low	low	low	low
Strategy	st	low	medium	medium	medium
	ss	medium	high	high	high
	us	high	high	very high	very high

The statistical results indicated that students did not perform significantly differently on the differently structured versions.

3.1.3.2.3 *Question 3*

Each part of this question was rated as follows.

Demand	Level of structure	a	b	c	d
Complexity	st	medium	medium	medium	medium
	ss	medium	medium	high	high
	us	high	high	very high	very high
Resources	st	medium	medium	medium	high
	ss	medium	medium	medium	high
	us	medium	high	high	high
Abstraction	st	low	low	low	low
	ss	low	low	low	low
	us	low	low	low	low
Strategy	st	medium	medium	medium	medium
	ss	high	medium	high	high
	us	high	high	very high	very high

Students did not perform significantly differently on the differently structured versions.

3.1.3.2.4 *Question 4*

Each part of Question 4 was rated as follows.

(see top of next page)

Demand	Level of structure	a	b	c	d
Complexity	st	medium	medium	medium	medium
	ss	high	high	high	high
	us	very high	very high	very high	high
Resources	st	low	low	medium	high
	ss	low	low	high	high
	us	low	low	very high	high
Abstraction	st	low	low	low	medium
	ss	low	low	low	medium
	us	low	low	low	medium
Strategy	st	low	medium	medium	medium
	ss	medium	high	medium	high
	us	high	high	high	very high

Statistically, the semi-structured version was most difficult in parts (a) and (c), whereas in part (b) there was no difference between the versions and in part (d) the difficulty of the question increased as structure decreased.

The examiner rated the unstructured version as being more demanding than the semi-structured on most dimensions for parts (a) and (c). The demands in the semi-structured version seem to be related to the word 'type' in part (a) and the misleading language in part (c), which were not demands intended by the examiner.

In summary we have seen no instances of more structured questions being perceived as more demanding by the examiners. In all questions and for all scales the level of demands either

remained the same or increased as the question became less structured. This reflected the statistics in some but not all cases.

To return to the list of demands affecting performance:

1) The use of an everyday term in a technical sense

2) The use of a difficult technical term

These two demands relate to the examiner's ratings on the scale of complexity in Question 1.

3) Specifying the organisation of the answer required

4) Providing cues to the content of an answer, changing the requirement from production to recognition

5) The use of different command words such as describe and explain, changing the nature of the answer required

6) Breaking down the question into sub-parts

The examiner's ratings on the scale of strategy in Question 1 correspond to these demands.

7) Asking students to come up with a theory rather than leading them into discovering it via a specific task

8) Misleading wording of questions resulting from an attempt at structuring

These demands from Question 4 were not identified by the examiner, and are not reflected in the ratings of the question. They arose as a side-effect of structuring the question.

3.1.3.3 Levels of demands in the responses

The examiners also described the responses they had marked in terms of the four dimensions on the scale of demands. These are detailed in the tables below.

3.1.3.3.1 *Question 1*

Demand	Effect on performance
Complexity	N/A
Resources	More structure gave more focus and direction for where to locate answer in resource, which led to better responses Less structured question required an element of interpretation.
Abstraction	Similar problems across all versions with concept of region and describing patterns
Strategy	Structure ensured that 'similarities' and 'differences' were given. Without the prompt to consider both, only differences were commented on. Had to devise own strategy which left their response incomplete and poorly organised.

The examiner rated parts (a) and (b) of the actual question as increasing in Complexity as they became less structured, whereas he thought the complexity scale was inapplicable to the responses. The comment on Resources contrasts with the ratings of demand in the question, in

which it was thought that demand remained the same in the three versions. As in the rating of the question, the level of abstraction remained the same in all versions, and more demanding strategies were being used in the responses to the less structured questions.

The statistical results indicate that students performed better on the structured version, and in terms of the demands seen by the examiner in the responses, this is related to the use of diffferent strategies in their answers.

3.1.3.3.2 *Question 2*

Demand	Effect on performance
Complexity	Less structured version was less daunting to lower ability students
Resources	Resources were equally well accessed across the versions.
Abstraction	More thoughtful answer, better linkages with increased structure. Less likely to evaluate success/failure with decreased structure.
Strategy	More detail and better organised responses with more structure.

The examiner's impression from the responses was that the structured version was more complex and daunting to the students, although the question itself had been rated as more demanding in the unstructured version. The statistical results indicated no change in performance across the versions, suggesting that the demands in the question and the demands displayed in the responses cancelled each other out.

The resources were not seen to change in the responses or the question. However the level of abstraction in the responses was thought to be greater for the structured version, despite there being no difference in the statistics or in the question ratings. For strategy, the question and responses were rated as increasing in demands as structure decreased, although the statistics showed no difference in performance.

3.1.3.3.3 *Question 3*

Demand	Effect on performance
Complexity	Links appeared to have been made across all versions
Resources	Better use of resources in structured version. Not so able to focus in on relevant parts of the resources in unstructured version
Abstraction	Better understanding and recognition of ideas in structured versions Less likely to recognise obvious ideas in unstructured
Strategy	Better use of case studies and more logical reasoning in structured Less detail, poorer organisation and diagrams in unstructured

The level of Complexity in the responses was thought to remain the same, despite the increased Complexity in the unstructured question. Students' performance was at the same level on all three versions.

The level of Abstraction in the question was seen as remaining at the same level of demand whereas the examiner thought this varied in the responses. This variation was not reflected in

the statistics. The variation in Strategy seen by the examiner in the question and the responses was again not reflected in the statistics, as performance on all three versions was similar.

3.1.3.3.4 *Question 4*

Demand	Effect on performance
Complexity	More linkages and more developed answers in structured version
Resources	Better use of place evidence and relative changes in structured version Without focus on resource less likely to use evidence
Abstraction	Less able to cope with ideas, and obvious patterns missed in unstructured version
Strategy	Sections omitted and more likely to choose an inappropriate study in unstructured version

The examiner rated the responses as differing on all four demands across the versions, which corresponds to the statistics showing that students performed better on the structured version than on the other two. However, the demands of resources and abstraction were seen to remain the same in most parts of the question, indicating that the difference in performance is due to complexity and strategy.

Returning again to the original list of demands,

1) The use of an everyday term in a technical sense

2) The use of a difficult technical term,

these demands, relating to complexity in the question, were not identified by the examiner in the response ratings.

3) Specifying the organisation of the answer required

4) Providing cues to the content of an answer, changing the requirement from production to recognition

5) The use of different command words such as describe and explain, changing the nature of the answer required

6) Breaking down the question into sub-parts

These demands were due to differences in the strategy indicated by the question in each version, and the examiner also identified differing strategies as occurring in the responses.

7) Asking students to come up with a theory rather than leading them into discovering it via a specific task

This demand was not identified by the examiner in the question. However, it was evident in the examiner's description of the use of resources in the responses to Question 4. He noted that students doing the structured question made better use of place evidence, that is when they were told to focus on things in common in the diagram they were able to come up with the theory, whereas they found this more difficult in the less structured question.

8) Misleading wording of questions resulting from an attempt at structuring

This demand was not identified by the examiner in the question or the responses, but as mentioned above, it arose as a side-effect of structuring the question.

3.1.3.4 Demands found using Repertory Grid technique

In order to discover further the examiner's ideas of what the demands were in the Geography GCSE questions, the Repertory Grid technique was used, and produced the following constructs.

	More demanding construct	Opposite
1	Structure not given	Structure provided
2	More abstract	More concrete
3	Explain situation	Describe
4	Resource new No resource provided Interpretation and analysis techniques must be developed	Resource familiar Resource provided Interpretation and analysis techniques familiar
5	Processing more complex issues	Processing simple information
6	Recall of information necessary	No recall of information necessary
7	Links less obvious	Links provided
8	Provide and analyse evidence	Provide evidence only
9	Remote / larger scale and less familiar subject	Local / small scale and more familiar subject
10	Prediction required	Current issues Present

Some of these constructs are very similar to the demands detailed in the Geography GCSE version of Edwards et al.'s (1985) scale. Each sub-question, or chunk, of Question 1 and Question 3 was then rated on a five point scale according to how demanding it was in terms of each of the constructs. The ratings were analysed using principle components analysis, and three factors emerged, suggesting that the constructs were based on three different types of demand. The factor weightings are detailed in the next table.

		Factor 1	Factor 2	Factor 3
1	**Structure**		.98	
2	**Abstraction**	.55		.64
3	**Explanation**	.95		
4	**Resource**	.93		
5	**Complexity**	.97		
6	**Recall**	.96		
7	**Linking**		.97	
8	**Analysis**	.96		
9	**Familiarity**			.88
10	**Prediction**	.34		.70

The demands of explanation, complexity, resources, recall and analysis all emerged strongly in one factor. Structure and links emerged as a second factor, and familiarity, abstraction and prediction as a third. It seems that the examiner was focusing on three types of demand, one to do with how complex the question and resource were and how much recall and analysis was needed; a second to do with whether or not students had to make their own links between points; and a third to do with how familiar and concrete the concept in the question was.

The ratings were also analysed using an ANOVA for each construct. The ratings on constructs 1 and 7, that is whether the examiner thought structuring was given, and whether links were given, were significantly affected by the level of structure in the questions. For the other constructs there was no significant effect of level of structure.

This factor analysis suggests that the demands identified by the examiner using the Repertory Grid technique fell into three clear categories, which can be labelled structure, familiarity and complexity. Familiarity is closely related to the dimension of abstraction in the scale of demands, and complexity is present in both scales. Resources, which is a dimension in the scale of demands is now amalgamated with complexity in the constructs. Strategy, the fourth dimension in the scale, does not appear as a construct, but could be seen to be related to the structure and links that emerged as one factor.

3.1.4 *Students' perceptions of demands*

Two pairs of students were interviewed while the rest of the group were doing the exam. They were each asked first to work through one version of Question 1 on their own, and then to talk about what they thought about the question, and what they thought of the other two versions. The themes described below emerged from analysing the students' remarks during the interviews.

- **Specificity**: The extent to which the content and process of answering were prescribed.

  Students preferred the semi-structured version as there was more flexibility in the response to this than to the structured version. Less structured questions allowed them

more opportunity to show their knowledge. However, they wanted some guidance on what was expected of them, without losing the flexibility. Too much information was restricting, but too little made a question vague, with more possibility for giving irrelevant answers. In some cases the structured version was perceived as expressing more clearly what the examiner wanted, for example for a four mark question that asked for similarities and differences students knew to point out two of each, whereas if the question had just asked them to describe the patterns they would have concentrated on either similarities or differences, not getting full marks. A balance is needed between telling them exactly what is wanted and allowing them to express what they know.

- **Prompts**: Pointers giving clues to relevant content.

  For example, Question 1(d) asked them to describe why a particular region was distinctive, and the structured version gave prompts to use physical and human features as headings. These prompts allowed them to know exactly what the examiner wanted. Without the prompts they may have concentrated for example only on human features.

- **Difficulty**:

  The structured version of Question 1 was thought to be the easiest, followed by the semi-structured, with the unstructured being the hardest. There was little to choose between the structured and semi-structured versions, with the structured questions often being seen as more friendly rather than easier.

- **Strategy**: How to approach the question and decide the content of the answer.

  The students mentioned the organisation of time, for example, leaving enough time for the last question if it had more marks than the others. Mark allocation was also used to organise responses in terms of how much to write for each question. One student mentioned using bullet points, with each point corresponding to one available mark. Some students felt they would write more for the structured version of Question 1, and that the unstructured version would cause the problem of not knowing what to include as it was vague.

- **Validity**: Whether the students perceived it as a fair question.

  They thought that it would be harder to get a level 3 (top mark) on an unstructured question because they would not necessarily have given all the detail required, and that the structured version should be worth fewer marks because it was easier. It was felt that the same mark scheme should not be used for the differently structured versions. For Question 1 in particular, some of the language in the unstructured version seemed to be easier than in the semi-structured version, for example 'accurate location' versus 'co-ordinates' (although 'location' did in fact cause more difficulty). The words in the

semi-structured version seemed more geographical, and although this made it hard, they also realised it would test their knowledge better.

- **Resources**:

  The labelling in the satellite picture was in the centre of the squares, whereas on a map two-digit numbers are normally used, and this caused some confusion. The key on the satellite picture was thought to be unclear, and should have included actual colours rather than colour names. The colours on the picture were not easy to distinguish. Also, the photograph did not have a compass, so for the question that referred to 'West of the Rhine' students had to assume that the top of the picture was North. In contrast, Figure 2, the cartoon, was seen as a welcome relief.

- **Language**:

  In the use of technical language, the semi-structured version was thought to be the best as it was a compromise between the other two, using some geographical words without the need to guess what is required. The use of the word 'location' in the unstructured version of Question 1(a) caused difficulty as the students did not know whether this meant give the co-ordinates or describe the area.

We can now return again to the list of demands:

1) The use of an everyday term in a technical sense

Students identified this demand as occurring in Question 1(a).

2) The use of a difficult technical term

This demand was not identified by the students. It occurred with the use of the word 'confluence' in the semi-structured version, but students felt that this version did not contain problematic words.

3) Specifying the organisation of the answer required

This demand was identified by the students as occurring in Question 1(b) where they said that with a less structured question they would have concentrated on similarities or differences and not both. This did in fact occur according to the examiner's ratings of the responses.

4) Providing cues to the content of an answer, changing the requirement from production to recognition

The students mentioned that the use of prompts in a question helped them to know what to write about, and this was reflected in the statistics and in the examiner's ratings on the scale of strategy.

5) The use of different command words such as describe and explain, changing the nature of the answer required

6) Breaking down the question into sub-parts

7) Asking students to come up with a theory rather than leading them into discovering it via a specific task

8) Misleading wording of questions resulting from an attempt at structuring

These last four demands were not mentioned by the students.

3.1.5 *The effects of structure on demands in GCSE Geography*

To summarise, the effects of different types of structuring on the questions in Geography GCSE were as follows:

3.1.5.1 Questions 1 and 4 - Structure caused a difference in performance

1) *The use of an everyday term in a technical sense*

Question 1a(i): The unstructured version contained the word 'location' used in a technical geographical sense, and some students misunderstood this and interpreted it as it is used in everyday language, referring to the type of area the town was in rather than simply giving co-ordinates as required. The use of this term made the question more geographically valid as well as more demanding.

2) *The use of a difficult technical term*

Question 1a(ii): The use of the technical geographical term 'confluence' clearly caused difficulty with some students not knowing its meaning. This is backed up by performance on the unstructured version which was very similar to performance on the structured version. The unstructured question asked students to suggest one reason for the growth of Koblenz on this 'site'. The word 'confluence' was not used, and students had little difficulty. Although structuring the question with the use of the word confluence made it more specific, and perhaps geographically more valid, it also made the question more difficult.

3) *Specifying the organisation of the answer required*

Question 1(b): Students attempting the semi-structured and unstructured versions concentrated only on differences and ignored similarities. The way students organised their answers was therefore affected by the structuring of the question.

4) *Providing cues to the content of an answer, changing the requirement from production to recognition*

Question 1(c): The structured version asked for a description of what the figure is predicting in c(i), and then in c(ii) asked students to suggest how two out of three given factors might explain the answer. In the semi-structured version c(ii) asked them to suggest two reasons such as economic and social factors. The unstructured version asked students to explain and give two reasons, without any further suggestions. Students were significantly helped by the prompts in the structured version, and the lack of prompts in the less structured versions caused difficulty. The prompts enabled students merely to recognise and select the information

to use in their answer, whereas in the less structured versions they had to produce their own answer to be in line with the examiner's expectations.

5) *The use of different command words such as describe and explain, changing the nature of the answer required*

In Question 1(c) students were asked to 'describe' and 'suggest reasons' in the structured and semi-structured versions, whereas in the unstructured version they were asked to 'explain' and this was more challenging.

6) *Breaking down task into sub-parts*

Question 4(d): The structured version was significantly less difficult than the other two. It asked students to choose a transport improvement they had studied and then gave three possibilities. The question was then divided into three parts, (i) asking them to name and locate it, (ii) asking them to describe the advantages it has brought and (iii) to comment on the conflicts. The semi-structured version asked students to name and locate a transport development they had studied and then went on to become a two-part question asking for (i) advantages it has brought to the area, and (ii) conflicts it has caused. It seems that the break down and the suggestions given in the structured version were a significant help to the students.

7) *Asking students to come up with a theory rather than leading them into discovering it via a specific task*

Question 4(a): The structured version asked students to study a figure, name two areas of employment decline and suggest what they have in common. The semi-structured version asked students to describe which type of area is experiencing employment decline. The students clearly had difficulty in working out which type of area they should refer to without the help given in the structured version which suggested that they look for things in common between two areas. The unstructured version asked students to describe the pattern of decline and was also easier than the semi-structured version. This indicates that it was the use of the phrase 'type of area' that caused the difficulty. In the semi-structured version, students were required to generate a theory of which type of area experiences decline, and were not given any help with doing so. In the structured version they were helped by being told to look at things in common, and in the unstructured version they were told to describe the patterns, and this process of describing resulted in an understanding of the theory. However, when required to generate a theory from scratch students found this question too difficult.

8) *Misleading wording of questions resulting from an attempt at structuring*

Question 4(c): The structured version was significantly less difficult than the other two, but again it was the semi-structured version that was the most difficult. The structured version asked how two out of three possible headings explain the location of employment growth. The semi-structured version asked students to explain how two factors, such as communications,

may explain the location of employment growth. This was confusing as it asked for two factors and then gave one factor, communication, as a prompt, leading students to concentrate entirely on communication in their answer, which excluded them from gaining high level marks. The unstructured version asked them simply to give two reasons for the pattern of growth. The semi-structured version in this case was confusing, and illustrates the dangers of imposing structure.

It should be noted that in Question 4 it was almost always the semi-structured question that was the most demanding. This was largely due to side-effects of imposing structure that the examiner had not intended.

3.1.5.2 *Questions 2 and 3 - Structure caused no difference in performance*

Questions 2 and 3 showed no significant effects of the structuring of the question on the performance of students. These questions can be considered in the light of the list of demands found in those questions where there was an effect of structuring.

4) Providing cues to the content of an answer, changing the requirement from production to recognition

Questions 2 and 3 contained some prompts in the structured and semi-structured versions but these did not help the students significantly. The unstructured version of Question 2(a) pointed the students straight to the graphs, which had the equivalent effect to giving them prompts.

5) The use of different command words such as 'describe' and 'explain', changing the nature of the answer required.

Question 2(b): This used the word 'explain' in the structured version and 'suggest' in the other two, which made the structured version difficult.

7) Misleading wording of questions resulting from an attempt at structuring.

The structured version of Question 2(d) asked students to 'comment on the success and/or failure' which was wordier than the other versions asking for a 'comment upon its outcome'.

In Question 3(c)&(d) the unstructured version was less wordy than the semi-structured, making it a slightly easier question.

These effects of structuring combined to make the three versions of Questions 2 and 3 of similar difficulty for the students.

3.2 Geography A Level

3.2.1 *Structuring the questions*

The questions used in the three versions of the Geography A Level papers were based on Question 1 in Paper 1 and Question 1 in Paper 2 of the Summer 1996 UCLES Geography (9050) examination. The examiner re-wrote the original A Level questions at three different levels of structure. As there was only one question on each paper, each question had two differently structured parts (see Appendix 1 for questions and mark schemes).

3.2.1.1 Question 1

Parts a(i), a(ii) and (b) were at one level of structure and part (c) was at a different level of structure. The examiner varied structure in a number of different ways. Structuring the process of answering the question was achieved by varying resources and layout. Resources were varied in the three versions, with part (b) of the structured and semi-structured questions including a diagram. For part (c) the structured version was broken down into two distinct parts.

The structured versions of both parts of the question also included prompts, and the unstructured version was made more abstract in part a(i) as follows.

Students were asked:

What terms are defined by the following descriptions?

In the structured version one of the descriptions was:

the ratio of river channel cross sectional area to wetted perimeter

In the semi-structured version the description was:

$$\frac{\textit{river channel cross sectional area}}{\textit{wetted perimeter}}$$

In the unstructured version the description was:

A/P where A is river channel cross sectional area
P is wetted perimeter

Wording was also varied, with the less structured versions containing more geographical terms.

Command words were changed, for example in part (c) where the structured version asked:

(i) Outline how the following can cause flooding:
precipitation events:
catchment conditions;
human activity;
(ii) How successful are methods used to predict floods?

Whereas the semi-structured version asked students to

Outline three causes of flooding and evaluate the success of methods used to predict floods.

The unstructured version used a different command word again:

Discuss how the success of flood prediction depends upon understanding the causes of flooding.

The use of words such as 'outline', 'evaluate', and 'discuss' has an influence on the type of response required, and the use of technical geographical terms can cause difficulty for the students. The use of prompts in the structured part (i) also changes the nature of the task.

3.2.1.1.1 *Question 2*

In this question parts (a) and (b) were at one level of structure and part (c) at another. Again the question was structured in a number of different ways. The structured versions were broken down into sub-parts in order to structure the process of answering the question.

Again, prompts were used in the structured version in part c(i) and notably in part (b) which was broken down into three sub-parts, each giving considerable help to the student. Part (b) also included different command words in the three versions, making the structured question quite different from the other two.

All three versions of part (b) began with telling students to study a figure.

Part (b) in the structured version then read:

State, giving your reasons, and evidence from the figure, whether you agree or disagree with each of the following:

(i) countries with higher rates of natural increase tend to have higher life expectancy;

(ii) countries with higher rates of natural increase tend to have lower infant mortality rates;

(iii) GNP per head appears not to be related to the demographic indices.

The types of relationships to look at are therefore given in the structured version. However, it is also an exercise in deciding whether these statements are valid or misleading.

The semi-structured version reads:

For the countries shown in Fig. 1, describe and explain three relationships which appear to exist between the demographic and economic indices.

Here students are told the sort of thing to look for but not at the level of detail that exists in the structured version.

The unstructured version reads:

For the countries shown in Fig. 1, describe and explain three relationships which appear to exist between indicators.

Here the students are given no guidance as to the sorts of relationships they should be looking at in the unstructured version. Also, the examiners may be expecting something different from students when the words 'describe and explain' are used and when the words 'state giving your reasons' are used.

A further method of structuring that may have affected the content of the question is the use of technical terms. Part (a) in the unstructured version used the term 'negative population growth', similar to the question in the original A Level paper, and this may have caused difficulty for students who could have answered the question if the term had been explained as in the structured question where it is referred to as ' death rate exceeding birth rate'.

In the next section the effect of this structuring on the performance of students on these questions is discussed.

3.2.2 *The effects of structure on difficulty*

3.2.2.1 Distribution of students across versions

The distributions of students' forecast grades across the six versions did not vary. The table below shows the numbers of students at each predicted grade who took each version. The six groups can be considered equivalent, since the variation between group means is not stastistically signficant.

	Paper 1			Paper 2		
Predicted Grade	Version A	Version B	Version C	Version A	Version B	Version C
A	13	5	17	16	5	16
B	17	14	16	18	14	16
C	10	18	14	11	20	14
D	9	16	13	8	15	15
E	6	9	6	7	9	6
N	0	0	0	0	0	0
U	0	0	1	0	0	1
Total	55	62	67	60	63	68
Mean Grade	5.4	4.8	5.3	5.5	4.9	5.3

3.2.2.1.1 *Question 1*

This table gives the mean marks of students on Question 1 for each version.

	Structured	Semi-structured	Unstructured
Q1(a&b)			
Mean mark	8.4	6.7	6.6
Maximum mark	13	13	13
Q1(c)			
Mean mark	6.7	7.0	5.7
Maximum mark	12	12	12

Performance on the unstructured version was below performance on the other two versions in both parts of this question. However, in (a) and (b) semi-structured performance was similar to unstructured, whereas in part (c) performance on the semi-structured version was closer to the structured, and in fact better than on the structured version.

A more detailed breakdown of how the methods of structuring affected each part of the question can be seen in the table below.

Q 1	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
a(i)	equal			$F_{(2,185)} = 1.5$, n.s.	Written definitions (st) vs Symbolic (ss,us)	None
a(ii)	st	ss	us	$F_{(2,185)} = 15.3$, $p<0.005$	'time of year'&'type of rock' (st) vs 'time and geology' (us)	Abstract and technical terms
b	st	us	ss	$F_{(2,185)} = 4.4$, $p<0.05$	Given names of subsystems (st, us) vs No direction (ss)	Prompts specifying organisation & content
c	ss	st	us	$F_{(2,185)} = 5.6$, $p<0.005$	'Outline' causes & 'evaluate' success (st, ss) vs 'Discuss' success and causes (us)	Command words Dividing into two tasks

3.2.2.1.2 *Question 2*

The next table gives the mean marks of students on Question 2 for each version.

	Structured	Semi-structured	Unstructured
Q2(a&b)			
Mean mark	8.4	8.3	6.9
Maximum mark	13	13	13
Q2(c)			
Mean mark	6.1	5.3	5.2
Maximum mark	12	12	12

In this question students performed similarly in the structured and semi-structured versions for parts (a) and (b), with poorer performance on the unstructured version. In part (c) performance on the semi-structured version was similar to the unstructured version. A more detailed breakdown of how the methods of structuring affected each part of the question are shown next.

Q 2	Easy	Mid	Hard	Significance	Structuring	Demands affecting performance
a	ss	st	us	$F_{(2,193)} = 20.0$, $p<0.005$	Birth and death rates (st, ss) vs 'negative population growth' (us)	Abstract technical term
b	equal			$F_{(2,185)} = 2.1$, n.s.	'State giving reasons' (st) vs Describe and explain (ss,us)	None
c	st	ss & us equal		$F_{(2,193)} = 5.9$, $p<0.005$	Given social factors to evaluate, then economic (st) vs All in one question with no prompts (ss,us)	Prompts specifying organisation & content Dividing into two tasks

The methods of structuring that produced demands that affected the level of performance in the questions are listed below.

1) The use of abstract and technical terms
2) The use of prompts specifying the organisation and content of an answer
3) Dividing a question into two separate tasks rather than asking students to discuss two things at once
4) The use of different command words, changing the nature of the task

3.2.3 ***Demands***

3.2.3.1 Scale of Demands

See section 3.1.3.1 for the scale of demands identified by the Geography examiners.

3.2.3.2 Examiner's ratings of level of demands

3.2.3.2.1 *Question 1*

The examiner rated each part of Question 1 on how demanding it was on each of the four dimensions on the scale of demands. The ratings are described in the table below.

Demand	Level of structure	a(i)	a(ii)	b	c
Complexity	st	medium	medium	medium	medium
	ss	medium	medium	medium	medium
	us	high	medium	medium	high
Resources	st	n/a	n/a	medium	n/a
	ss	n/a	n/a	high	n/a
	us	n/a	n/a	n/a	n/a
Abstraction	st	low	low	medium	medium
	ss	medium	medium	medium	medium
	us	high	high	medium	medium
Strategy	st	medium	medium	low	low
	ss	medium	medium	medium	medium
	us	medium	medium	medium	high

According to the statistical results, part a(i) yielded no significant effects of structure in the statistics reported above, whereas in part a(ii) students performed better on the structured version and worst on the unstructured, indicating that the change in demands relating to abstraction had an effect. The statistics showed that students performed better in part (b) on the structured than the semi-structured version and this may have been due to the level of guidance they were given in the use of the resource. For part (c), the statistics indicated that the unstructured version was more difficult than the other two, and according to the examiner's ratings, this is due to differences in complexity and strategy.

3.2.3.2.2 *Question 2*

Demand	Level of structure	a	b	c
Complexity	st	high	medium	medium
	ss	low	medium	medium
	us	high	high	high
Resources	st	n/a	medium	n/a
	ss	n/a	medium	n/a
	us	n/a	medium	n/a
Abstraction	st	high	medium	high
	ss	low	medium	high
	us	high	medium	medium
Strategy	st	low	low	low
	ss	medium	medium	medium
	us	high	high	high

The examiner rated part (a) as more demanding in the structured and unstructured versions than in the semi-structured version in terms of complexity and abstraction. The statistics also show that students doing the semi-structured version performed better than those doing the other two versions. In part (b) the task of stating whether you agree or disagree with the statements was thought to require a less demanding strategy than describing and explaining the relationships between indicators. However, no significant difference in performance in the three versions was indicated by the statistics. The examiner rated the structured and semi-structured versions of part (c) as more demanding than the unstructured in terms of abstraction,

because they referred to terms such as social and economic factors. The statistics showed that the structured version, containing prompts, was easier than the other two.

The demands identified by the examiners on the four scales have helped to clarify some of the reasons for the findings detailed in the results section above. To return to the list of methods of structuring that affect demands:

1) The use of abstract and technical terms

This corresponds to the examiner's ratings on the scale of abstraction. Abstract terms were thought to introduce demands.

2) The use of prompts specifying the organisation and content of an answer

3) Dividing a question into two separate tasks rather than asking students to discuss two things at once

4) The use of different command words, changing the nature of the task

These methods of structuring relate to the examiner's ratings on the scale of strategy

3.2.3.3 Levels of demands in the responses

The examiner also described the responses in terms of the same four dimensions on the scale of demands, based on an overview of the responses he had marked.

3.2.3.3.1 *Question 1*

Demand	Effect on performance			
	ai	aii	b	c
Complexity		Complex terms (us)	Direction set them on right track (st)	Extended range of answer (st)
Resources		Difficulty drawing diagram (ss)	Repeated diagram rather than discussing complex issues (ss)	
Abstraction	Literary definition (st) produced better responses than abstract formula (us)	Water table difficult (us)		Question on flood prediction demanding (st)
Strategy	Followed structure for strategy, so better organised (st)	Didn't use diagram (us)	Used more than one subsystem (ss)	Need to reverse question for suitable strategy (us)

Overall for this question, the examiner thought that students made links more easily when there was structure, and the direction in the use of resources helped the weaker students but not the more able students. Although the examiner thought the question became more demanding with less structure, this was not always reflected in the students' performance.

3.2.3.3.2 *Question 2*

Demand	Effect on performance		
	a	b	c
Complexity	Better responses with set problem (st)	Structuring helped organisation, although some ran out of time (st) Helped give scale to task (ss)	Better performance when question divided, good use of headings (st)
Resources		No self-generation (us) More expansive answers but repetitive (st)	
Abstraction	'Negative population growth' (us) seen as value term Migration (ss) more challenging than birth & death rates	Difficulty abstracting from graphs (us) Structuring inhibited abstract thought and complex ideas (st)	Abstraction reduced by country by country approach (us)
Strategy	Some students missed migration (ss)	Students only gave direct interpretations (us)	Country by country approach rather than social and economic (ss, us)

To reconsider the list of ways of structuring:

1) The use of abstract and technical terms

The examiner identified this as having caused differences in demands in the responses to the definitions in Question 1 a(i) and the technical terms in Question 2(a).

2) The use of prompts specifying the organisation and content of an answer

This was identified in the responses to Question 1(b) and (c) in which direction helped the students, and also in all parts of Question 2.

3) Dividing a question into two separate tasks rather than asking students to discuss two things at once

The responses to Question 1(c) were thought to differ according to whether or not this was in two separate parts, as were the responses to Question 2(c).

4) The use of different command words, changing the nature of the task

These were not mentioned by the examiner as having affected the responses.

3.2.3.4 Demands found using Repertory Grid technique

Each construct that emerged from A Level Geography is listed below, along with its opposite.

	More demanding construct	Opposite
1	Preparation - have to understand complex diagram. Can't be pre-learned	Straight into the question - no diagram or familiar diagram. Set steps taught
2	Vocabulary complex	Everyday language, no technical terms
3	Resource - complex/unfamiliar. Describe and explain a relationship	Describe without explanation for a simple resource
4	Focus - several different relationships to be identified	Focus on single message from data
5	Trigger word hints at uncertainty	Indication of single unambiguous answer
6	Variable range - measures on different scales	No specific variables, just description terms
7	Abstract conceptual processes, mathematical formula	Concrete visible processes
8	Broad statement to be explored	Divided into series of steps
9	Involving past and future, processes no longer active	Contemporary processes which can be seen in the field. Everyday common experience
10	Large scale, global, remote from experience	Small scale, matching experience
11	Command words - evaluate, explain	Describe, define

Some of these are very similar to the demands detailed in the Geography A Level version of Edwards et al.'s (1985) scale. Constructs 1 and 3 relate to the resources dimension of the scale of demands, and 7, 9 and 10 refer to abstraction. Numbers 4, 5, 6 and 8 may relate to the scale of complexity and 2 and 11 to strategy.

Each sub-question, or chunk of each question, was then rated on a five point scale according to how demanding it was in terms of each of the constructs. The ratings were analysed using principle components analysis, and four factors emerged, suggesting that the constructs were based on four different types of demand. One factor consisted of constructs relating to resources, another included constructs relating to structure, such as number 8, the third consisted of constructs such as abstraction and time, and the fourth included most of the other constructs and could be labelled 'complexity'. The factor weightings are detailed in the next table.

		Factor 1	Factor 2	Factor 3	Factor 4
1	**Preparation**	.30	.85		
2	**Vocabulary**	-.32		.55	.60
3	**Resource**		.83		
4	**Focus**	.66			.53
5	**Uncertainty**	.73	.56		
6	**Variables**		.58	.37	
7	**Abstraction**			.86	
8	**Broad statement**				.86
9	**Past/future**			.83	
10	**Global**	.87			
11	**Command words**	.48			.49

The ratings made by the examiner on each construct were also analysed using an ANOVA for each construct. For constructs 2 and 8, that is vocabulary and broad statement vs series of steps, there was a significant effect of level of structure on the ratings given by the examiner. Questions of different structure were seen to be making different demands on students in terms of these constructs. For all of the other constructs there was no significant effect of level of structure on the ratings given by the examiner.

3.2.4 *Students' perceptions of demands*

Two pairs of students were interviewed after they had done the exam. They were each asked first to make notes on their own of how they would approach one version of the question, and then to talk about what they thought about the question, and what they thought of the other two versions. Six themes emerged from analysing the students' remarks during the interviews.

- **Specificity**: The extent to which the content and process of answering the question were prescribed

  Students were concerned about knowing what the examiner expected, and the structured version, being more broken down, allowed them to see this more clearly. There were also negative comments about structure. They felt that too much structure was constraining and restricting, and did not allow them to use their own points or to demonstrate their understanding and knowledge. The other problem with a highly structured question was that when it was very specific students either knew it or did not, and if they had not revised that topic they may not have been able to answer it at all, whereas in a less structured version they could 'put in anything at all' or 'tackle it your way'.

  The structured versions sometimes did not ask for conclusions, which students were keen to give. A two-part structured question could result in them doing more work

than they would for a single essay as they would tend to write an essay-type answer for both parts even when this was not required. This may also have resulted in time problems for some students.

- **Prompts**: Pointers giving clues to relevant content.

  Prompts were seen as useful in stopping students from going off at tangents, but could also leave them with nothing left to say if too much was given in the question. Sometimes prompts can be too obvious and therefore confusing as students may feel they have missed something.

- **Difficulty**:

  One student thought that the unstructured version of Question 2 was easiest, especially for those who knew the answer, as it asked for a straightforward definition in part (a). This reflected the overall view of these students, who tended to prefer the unstructured versions, perhaps with the addition of a few prompts.

- **Strategy**: How to approach the question and decide the content of the answer.

  In the unstructured version students had to work out a strategy before starting the Geography in the question, and this took time. Without a given structure they had to create and impose their own structure, and in Question 2 these students approached the unstructured question in much the same way as the structured version was presented. However in Question 1, one student who was interviewed had tried to impose structure in part (c) of the unstructured version but had not imposed as much structure as was in the other two versions. The unstructured versions were thought to involve pulling things together, whereas the structured versions involved recounting facts.

  The students also used the mark allocation to determine how much time to spend on a question and how much to write. In the unstructured version, where there are 5 marks for a definition, the students thought this was too many and they would include more than necessary in case they had missed the point. Where there were 8 marks overall for a three part structured question one student would use a couple of marks to draw it all together, even though the question does not ask for this.

  This student said:

  'These types of ambiguous questions mean there is less time for the Geography'. Overall they felt that structure was better at the opening of a question, to lead them in, and it could also be appropriate when extracting information from data given.

- **Expectations**: The kind of question students have been taught to tackle

  These students were well trained in how to answer essay questions and so preferred the unstructured versions because they knew what they were doing and what the examiner

wanted. These students admitted to finding it difficult to ignore their expectations and form an objective view of structured questions.

- **Validity**: Whether the students perceived it as a fair question.

  These students thought that it was unfair that the mark scheme was awarding the same thing in each version, as they felt that the versions were actually different questions asking for different answers. It was also mentioned that the unstructured questions sometimes seemed to be trying to 'trick' students by making them work out exactly what the question was about. Mark allocations needed to be given to all question parts, so that students knew how much the examiner wanted for each part. In Question 1(b) in the structured version there were 8 marks allocated and students felt that this was too many as it covered only a very small part of the syllabus.

  Also related to validity, the students felt that unstructured questions were more appropriate for A Level as they tested understanding and also application of knowledge.

- **Resources**:

  The students felt that there were too many diagrams in Question 2, and this confused them as trends were found on some of these but not on others. The fact that they had to turn a page to look at the resource was also a cause of difficulty. For Question 1 the version without the diagram was thought to be easier because marks may be available for reproducing the diagram as well as for an explanation of it.

- **Language**:

  The questions asking for explanations were harder than those asking for descriptions.

Looking back at the list of methods of structuring that affected performance, it can be seen that some of these demands were identified by the students:

1) The use of abstract and technical terms

This was not mentioned by the students in their comments about the questions.

2) The use of prompts specifying the organisation and content of an answer

The students found prompts useful but thought that they could also be restricting. When no specification was given they imposed their own organisation, which in Question 1 matched the structured version although in Question 2 it did not.

3) Dividing a question into two separate tasks rather than asking students to discuss two things at once

They felt that the questions that were broken down allowed them to see more clearly what the examiner was expecting.

4) The use of different command words, changing the nature of the task

The students felt that it was harder to give an 'explanation' than a 'description'.

3.2.5 *The effects of structure on demands in A Level Geography*

1) The use of abstract and technical terms

Question 1 a(i) asking for definitions of terms was more abstract in the semi-structured than the structured, and even more abstract in the unstructured version, although students performed equally well on all three versions. Question 1 a(ii) also used abstract terminology in the less structured versions, and students performed poorly on these. Question 2(a) was more abstract in the unstructured than in the semi-structured, also resulting in poorer performance on the unstructured version.

The use of abstract technical terms affected performance in two out of the three cases.

2) The use of prompts specifying the organisation and content of an answer

This clearly had an effect in Question 1(b) where students were given suggestions of which sub-systems to focus on in the structured version, and performed well. In the unstructured version they were also given suggestions although these were not so specific. However in the semi-structured version no suggestions were given at all and students performed poorly in comparison.

Question 2(b) also contained prompts in the structured version, and this kept students' answers on track and allowed them to be more expansive although they did not score higher marks. This may have been because the structuring here also resulted in repetitive answers. The prompts in Question 2(c) also helped students, allowing them to be more specific in their answers, and score higher marks in the structured version.

3) Dividing a question into two separate tasks rather than asking students to discuss two things at once

In Question 2(c) when the structured version was divided into two tasks, addressing first social and then economic factors, students scored higher than when the tasks were combined. Also in Question 1(c) students writing about causes and then about successes of prediction did better than those tackling the unstructured version in which they had to discuss how prediction depends on causes.

Elsewhere where questions were divided students were also given prompts, making it difficult to determine the effects of dividing alone.

4) The use of different command words, changing the nature of the task

In Question 1(c) students were asked to 'outline' and 'evaluate' in the more structured versions, whereas on the unstructured they were asked to 'discuss' and this proved a more difficult question. The students themselves identified command words as changing the demands of a question.

Question 2(b) asked students to 'describe and explain' in the semi-structured and unstructured versions, whereas the structured version asked them to 'state giving reasons', although in this case it did not affect the performance of students on the three versions.

3.3 Combined summary for GCSE and A Level Geography

- Students performed significantly better on the structured version than the other two versions in GCSE Questions 1 and 4, and also in five out of the seven sub-questions in A Level.
- In GCSE Questions 2 and 3 students performed equally on all versions.
- The methods of structuring in GCSE affected performance by changing the cognitive demands as follows:

  1) The use of an everyday term in a technical sense

  2) The use of a difficult technical term

  3) Specifying the organisation of the answer required

  4) Providing cues to the content of an answer, changing the requirement from production to recognition

  5) The use of different command words such as describe and explain, changing the nature of the answer required

  6) Breaking down the task into sub-parts

  7) Asking students to come up with a theory rather than leading them into discovering it via a specific task

  8) Misleading wording of questions resulting from an attempt at structuring

- The methods of structuring in A Level affected performance by changing the cognitive demands in similar but fewer ways:

  1) The use of abstract and technical terms

  2) The use of prompts specifying the organisation and content of an answer

  3) Dividing a question into two separate tasks rather than asking students to discuss two things at once

  4) The use of different command words, changing the nature of the task

- Examiners rated the questions on four types of demand: complexity, resources, abstraction and strategy. The questions in which structure affected performance had different demands in terms of complexity and strategy in GCSE; and complexity, strategy and abstraction in A Level.

- Some of the cognitive demands introduced into questions were not intended by examiners, and these account for the cases in which semi-structured questions were more demanding than unstructured.
- The examiners identified different demands in the responses to the three versions, which corresponded more closely than the question ratings to the demands that affected performance.
- The Repertory Grid technique resulted in three constructs emerging in GCSE, relating to the demands of complexity, abstraction and structure/strategy; and four constructs emerging in A Level which were complexity, resources, abstraction and structure/strategy.
- Students perceived demands in the questions to be different in the three versions at both GCSE and A Level. They thought the structured questions showed more clearly what the examiner expected, although these questions were also restrictive and less flexible than the unstructured questions which allowed them more opportunity for demonstrating their knowledge.
- In conclusion, structuring in Geography involved changing cognitive demands in a number of different ways, which resulted in the more structured questions being easier in some but not all cases. A more structured question is therefore not necessarily an easier question in Geography.

4. CHEMISTRY

4.1 Chemistry GCSE

Two examiners wrote the papers used in the trial. There were four questions on paper 1 and five questions on paper 2. Most of these questions contained separate 'chunks' which were at different levels of structure. These 'chunks' did not necessarily have the same question number, since part of the structuring process often involved breaking the question down into extra parts, or reducing the number of parts. 'Chunk' therefore refers to blocks of the same content. The examiners' approach was to write a question that resembled a GCSE question in terms of structure (and content), and label this one the semi-structured (SS). They would then reduce the amount of structure to form the unstructured (US) version, and increase the amount of structure to form the structured (ST) version.

The data on the forecast grade of each candidate (see the table below) suggested that the ability distribution of the groups taking each version of the paper was approximately the same, although version C was taken by a slightly higher proportion of very able candidates.

	Paper 1			Paper 2		
Predicted Grade	Version A	Version B	Version C	Version A	Version B	Version C
A* - A*/A	13	9	17	14	10	17
A - A/B	13	18	8	10	17	7
B - B/C	12	11	12	13	10	12
C	6	5	6	5	6	7
D	0	1	0	1	2	0
E	0	0	1	0	0	1
Total	44	44	44	43	45	44

However, since there was a mixture of types of question on each version, any differences should have cancelled out overall.

4.1.1 *The effects of structure on difficulty*

In almost every chunk that was different across the versions, the US version was harder (i.e. had a lower mean mark) than the ST version. The mean mark of the SS version was usually in between the ST and US. The mean marks for each whole question are shown in the next table.

Paper 1	level of structure	mean	max mark	Paper 2	level of structure	mean	max mark
Q1	US	6.2	8	Q1	US	5.0	12
	SS	7.1	8		SS	7.2	12
	ST	7.9	8		ST	10.2	12
Q2	US	4.4	6	Q2	US	3.6	6
	SS	4.5	6		SS	4.0	6
	ST	5.6	6		ST	5.0	6
Q3	US	6.7	22	Q3	US	1.8	7
	SS	8.9	22		SS	2.4	7
	ST	12.3	21		ST	2.3	6
Q4	US	4.0	14	Q4	US	6.6	15
	SS	6.8	14		SS	7.0	15
	ST	10.4	14		ST	8.4	15
Total	US	20.0		Q5	US	2.1	10
	SS	27.3			SS	4.0	10
	ST	35.4			ST	4.3	10
				Total	US	19.2	
					SS	24.7	
					ST	30.3	

The overall size of the effect of structure was large - by aggregating the mean scores on the same version over the full 100 marks for paper 1 and 2, and assuming that the SS version represents the GCSE standard, we can say that adding structure could give candidates an extra 14 marks out of 100, and removing structure could reduce the score by 13 marks out of 100. Plots of mean marks for each 'chunk' are in appendix 2.

Across the two papers there were 25 different question 'chunks' of the same or similar content and number of marks available. The manipulations of structure which did and did not affect performance are described below, in terms of the demands made. It should be noted that it is almost never possible to vary a single aspect of structure.

1. 'Pure' structuring. This is the only form of structuring which had no significant effect on performance in GCSE Chemistry. This is where the difference involved asking for essentially the same information in a single sentence as opposed to several discrete parts. This was observed on 4 occasions, for example P2 Question 4A where the US version was 'Describe how the number of protons, neutrons and electrons change across the period from sodium to argon'. [4 marks]. The ST version asked the question three times, for protons, electrons and neutrons, for 1,1, and 2 marks respectively. A second example (P2 Question 5B) involved a calculation: the SS version was 'Calculate the total energy required to break the bonds in 1 mole of hydrazine and 1 mole of oxygen' [2 marks]. The ST version was d) 'Calculate the energy required to break 1 mole of hydrazine molecules into nitrogen and hydrogen atoms.' e) 'What is the energy required to break 1 mole of

oxygen molecules into oxygen atoms?' [1 mark each]. The demands that vary here are processing a longer sentence, and judging how the marks will be allocated. The latter was important to the candidates (see section on candidates' perception of demands).

2. Information. An example was P1 Question 1a where in the SS version candidates had to fill in the number of electrons in an empty table, whereas in the ST version some of the blanks had been filled in for them. Obviously in general this form of structuring reduces the demand of recall; in this case it also allowed the ST candidates to work out the answer from the information provided. Performance was better in the ST version.

3. Specificity. This was the most common form of structuring and it almost always included 'pure' structuring - i.e. breaking a question into parts. However, it differed in that different demands were made too. There were 5 cases where this happened, and the extra specificity always improved performance. An example is P1 Question 3F where the US and SS versions differed only in 'pure' structuring (see the appendix), essentially asking how to prepare lead II chloride from lead II nitrate. Performance was not significantly different. However, the ST version asked a couple of specific questions: 'What would you see when hot solutions ... are mixed and allowed to cool to room temperature?' and 'How could lead II chloride crystals be removed from the solution?'. This version was significantly easier. It removed the need to deduce from the data (or recall from memory) the appropriate method.
   A second example is P2 Question 4C, where the ST and SS versions differed only in 'pure' structure, and not significantly in performance. (See appendix 2 for full questions). The US version, which was significantly more difficult, was 'Explain how sodium and chlorine atoms combine to form a sodium chloride lattice.' [5 marks]. The SS version differed from this in including two specific questions which the US candidates were expected to answer without being specifically asked, namely: 'Show the formulae of the particles formed' and 'What forces hold the particles together in a sodium chloride lattice?' The difference in demand is that in the US version candidates have to know what the mark scheme will contain (from teaching and experience) - unless it is the case that the correct answer would obviously and unambiguously contain the points requested in the SS and ST versions.

4. Language. In a Chemistry exam, it is arguable that it is less valid to test linguistic skills than in other subject areas (e.g. History). Here 'linguistic skills' could be defined as the ability to express knowledge and understanding by constructing a comprehensible English sentence or paragraph. Often the US version of a question would require the candidate to fill 3 or 4 lines with writing. An example is P1 Question 1A where the SS version asked candidates to 'Complete the table by writing in the number of electrons in each orbit for an atom of each element'. This was significantly easier than the US version: 'Describe the arrangement of electrons in orbit around the nucleus of an atom of potassium.' (This example illustrates how difficult it is to vary only one feature of structure. Candidates are

asked for less information in the US version (lithium and sodium are not required), but the layout of the table in the SS version with column headings for '1st orbit', '2nd orbit' etc. gives extra information. Frequently when structuring removed the linguistic demands it altered the cognitive demands in other ways. For example in P1 Question 2, the ST version asked candidates 'Which of diagrams A to E could be i) air, ii) sulphur dioxide etc. The SS version, which was harder, was 'Suggest why C represents a mixture of gases.' etc. In this version the answer is supplied and the demand is in demonstrating understanding by explaining. In the ST version the answer is not supplied, and the candidate's understanding is inferred from their selecting the correct diagram.

5. Response Format. When the response format involved choosing a letter, ticking a box, putting a ring round a formula etc, performance was always better. This happened in the ST version of questions on 6 occasions. It is hard to say whether this was due to the lesser demand of recognising a correct answer than producing it, or due to the chance of getting marks for guessing. An example was P1 Question 1D, where the SS version asked candidates to suggest suitable values for the pH of hydrochloric acid and the neutral solution, and the ST version asked candidates to select two suitable values from a choice of four.

   Another form of structuring was to leave blanks in a sentence to be filled in by candidates choosing words from a list (known as the 'cloze' item type in language testing). This happened in the ST version of two questions and performance was significantly better on both. This form of structuring reduces the linguistic demand of constructing a sentence, but it allows candidates to select answers based on grammatical considerations - i.e. which word would fit, so does not rule out linguistic skills completely. It also provides information (reduces the demand of recall), increases the specificity, and allows guessing, so it is hard to assess the effect of this form of structure!

6. Strategy. This has appeared in examples previously discussed, e.g. where extra specificity or information gave clues to the correct strategy. The most concrete example was in a question involving a calculation - P1 Question 3D. The ST and SS versions asked identical questions, although the ST version came out with a slightly higher mean mark. The US mean mark was very low (0.22 out of 4). The structuring here took the form of splitting the calculation down into three steps in the SS and ST versions, and giving a hint in step 2: 'Use the equation to calculate...' This is a good example of the effect of structuring because the US version asked almost the same question as the third step of the other two versions, but without any lead-up questions.

In summary, it was possible to identify five features of structure which affected performance, and one which did not. However, it was very rarely the case that it was possible to vary a single aspect of structure, and variations in structure affected the demand in interacting and overlapping ways.

4.1.2 *Examiners' ratings of demands*

For each of the 25 question chunks in each version (the question papers and table of chunks are in appendix 2) we had the following data:

- examiners' ratings of the demands of the question on the dimensions of Complexity, Resources, Abstraction and Strategy - C, R, A and S. For a definition of these terms as they were used in Chemistry, see the table below. The scale was conceptually a 5-point one, with points 2 and 4 defined by descriptions.
- examiners' rating of the students' responses to the questions, using the same scale and dimensions.

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

There was thus a matrix of 600 ratings (25 'chunks', 3 versions of each chunk, 4 dimensions of the scale, and in each case a rating of the demands made by the question and the demands evident in the response).

The discussion below is based on comparisons made by breaking this matrix up in different ways.

The raters were reluctant to go outside the boundaries defined for them - i.e. nearly all ratings were '2', '3' or '4'. There were no ratings of '5', and the only ratings of '1' were given to the responses of a (structured) question which required candidates to put a tick against two substances produced from ammonia. In total, there were 25 separate question chunks across the two papers. In 18 of these, the structured version of the question was given a rating of '2'

(the lower defined point on the scale) on all 4 dimensions of the scale, both for the demands of the question and the demands evidenced in the responses. This suggests that the structured questions in this study, in the main, contained as few demands as possible, in all respects.

This data showing how the scale was used for each level of structure is summarised in the next table. For each level of structure, there were 25 question chunks, each rated on 4 dimensions in terms of level of demands required, and level of demands shown in the response. This means there were 25x4x2 = 200 ratings made for each level of structure.

	Rating			
Version	1	2	3	4
Structured (ST)	2	184	6	8
Semi-structured (SS)	0	160	26	14
Unstructured (US)	0	127	20	53

In general, the US version was rated as most different (more demanding), obtaining a rating greater than '2' on 73 out of 200 possible ratings. The vast majority of all ratings were in the '2' category, effectively at the bottom of the scale of demands as it was used here.

4.1.2.1 Differences between the ratings of demands required and demands observed in the response

In the 25 chunks, 3 levels of structure, and 4 dimensions, there were thus 25x3x4 = 300 opportunities to discover a difference between 'intended' demands and 'achieved' demands. Such a difference was only recorded in 25 instances - 12 of these on the same question chunk (where the structure had not been altered across each version). The general impression is that it is not really possible (in Chemistry) to separate the demands the question makes from the demands the candidates are able to show in their answer. This suggests that if the goal of testing in Chemistry is to enable candidates to produce evidence of 'higher order skills', such as planning a strategy, or organising / selecting data (which were the higher points of the dimensions S and R on our scale) then it is necessary to put less structure in the questions. This might seem self-evident, but it contrasts with the idea of 'levels of response' marking which is used in some subject areas.

4.1.2.2 Differences between the different dimensions on the scale of demands

The number of times each dimension was rated at each score category on the scale is shown in the following table:

Dimension	Rating			
	1	2	3	4
Complexity	0	100	20	30
Resources	1	142	7	0
Abstraction	0	121	9	20
Strategy	1	108	16	25

'Resources' was only given a different rating from '2' on 8 occasions out of 150 (25x3x2). Obviously this dimension does not play a significant part in these experts' concept of demands, at least on the questions involved here.

The raters were able to make more distinctions among the chunks in terms of 'Abstraction', but significantly there was virtually no variation across the levels of structure on this dimension. This suggests that demands due to level of abstraction are more topic-related than structure-related.

It was in terms of 'Complexity' and 'Strategy' that the most distinctions were made. Complexity relates mainly to the requirement to make links between ideas, and Strategy relates mainly to the requirement to devise and monitor a strategy. However, the raters did not often rate questions differently on these two dimensions. In the US version (which had the most ratings higher up the scale than '2') 18 of the 25 chunks were rated the same on Complexity and Strategy.

The analysis of the different types of structuring in section 4.1.1 showed that it was only very rarely possible to vary a single aspect of structuring and hence the demands the question made. This has also been found in the experts' rating of the demands of the questions and answers using this scale. Only 2 of the 4 dimensions, Complexity and Strategy, were related to the level of structure, and the ratings on each of these dimensions were usually the same for any given chunk.

4.1.3 *Students' perceptions of demands*

Pairs of candidates were interviewed in each subject using the procedure described in chapter 2. The views of two candidates for the GCSE Chemistry could be summarised as follows:

- Structured questions were generally considered easier and unstructured questions more difficult. This was sometimes based on the look of a question - one quote for a US question (P1 Question 2) was "The appearance is scary, but the question is not actually that hard."
- One advantage of a structured question was that it was clear what was expected. Unstructured questions could be vague and candidates unsure of what would get the marks. A second advantage (in their eyes) was that a simple response format removed the need to construct a sentence: "But the (P1 Question 1 ST) asks for numbers so people don't have to understand your written English, it's almost a multiple choice." They also welcomed the way structuring could organise the response, for example on P1 Question 4 "The (ST) gives you the questions you should ask yourself if you're doing the (US) version."

- A disadvantage of structure was that sometimes there was not enough space for the candidate to say all they wanted to say. They preferred the US version of P2 Question 1 because it was "good to do a description on", and felt restricted on this question by the SS and ST versions.
- An advantage of having less structure was that the greater generality meant that candidates could write what they knew and hope to pick up some marks, whereas if they couldn't answer a specific (ST) question, they were stuck. They were aware that with US questions there were often more mark points available in the mark scheme than the total for the question, giving a bit of flexibility and more chance of gaining marks.

Overall, their preference was for something like the semi-structured version (which is what they are used to). "I'd prefer something between 'no questions' and 'walking you through'".

It is clear that the candidates' comments relate to some of the features of structure described in the previous section. Comments about the appearance of the question and being able to see how many marks are available relate to 'pure' structure - although this was not found to have any effect on performance it clearly affects the 'morale' of the candidates as they attempt questions, which may be of non-negligible importance over a complete examination. Candidates appreciated questions where it was clear what the marks would be awarded for - in other words the more specific questions. They acknowledged the difficulty of writing sentences (language) and organising their answer (strategy). However, they also commented on aspects of structure that they disliked, or felt restricted by. These were high ability candidates, so this is a reminder that different styles of questions may be appropriate to allow all candidates to show what they can do.

4.1.4 *The effects of structure on demands in GCSE Chemistry*

1. Making a question less structured makes it more difficult, in terms of lowering the mean mark. The effect is noticeable even at an individual question level, and accumulated over a whole paper it could make a difference of at least 12 marks out of 100. Increasing the amount of structure could increase the mean mark by the same amount.
2. Six aspects of structuring used can be identified, five of which affect performance. However the 'purest' form of structuring - breaking the question into several parts without providing or requesting any extra information - has no significant effect on performance. It is very rarely possible to vary a single aspect of structure in a question. There are complex interactions between the types of demand introduced and removed by the other forms of structuring - which one is more appropriate is a question of validity to be decided by judgement.
3. In GCSE Chemistry, there is no significant difference between the demands required by the question and the demands candidates are able to show in their response. The implication is

that to test higher-order skills such as linking ideas and planning a strategy, less structured questions need to be set.

4. By comparing experts' ratings of different questions for different conceivable aspects of demands, it is clear that for the GCSE Chemistry questions in this study, manipulations of structure have altered the demands on the dimensions of Complexity and Strategy, but not Abstraction or Resources.

5. Complexity relates mainly to the need to make links between ideas and to demonstrate understanding rather than straight recall. Strategy relates mainly to the requirement to devise and monitor a strategy. In the majority of cases the manipulation of structure affected both these dimensions. The questions here were not considered to require the higher levels of demand on the dimension of Resources - generating information and discriminating relevant and irrelevant information. The level of Abstraction occasionally varied between questions, but not across the levels of structure, implying that this 'demand' is topic-related rather than structure-related.

6. From the candidates' point of view structured questions are less intimidating, and generally easier. They are clear about what they are expected to produce in response. They recognise that some questions are more appropriate in an unstructured form, and that for a candidate on top of the material it might be easier to gain marks on this type of question.

4.2 Chemistry A Level

One examiner wrote the questions to be used in the trial by adapting some of the questions that had appeared on the June 1996 UCLES Chemistry A Level examination (syllabus 9254). In the trial papers there were three questions on Paper 1 and four questions on Paper 2. Most of these questions contained separate 'chunks' which were at different levels of structure.

4.2.1 *The effects of structure on difficulty*

The forecast grades suggested that the 'ability' distribution of the groups taking each version was approximately the same with the possible exception of paper 2 where version B had slightly more candidates with a predicted 'B' and fewer with a predicted 'C' than the other two.

	Paper 1			Paper 2		
Predicted Grade	Version A	Version B	Version C	Version A	Version B	Version C
A	20	24	19	19	21	20
B	10	14	12	8	15	11
C	11	8	11	11	5	10
D	12	8	6	8	7	8
E	6	3	8	5	3	7
N	4	4	2	4	4	2
U	1	2	1	0	1	1
Total	64	63	59	55	56	59

Again there was a mixture of types of question on each version, so any differences should have cancelled out overall.

The mean mark on each whole question is shown in the next table. Data for each 'chunk' are in appendix 2.

Paper 1	level of structure	mean	max mark	Paper 2	level of structure	mean	max mark
Q1	US	3.2	12	Q1	US	3.2	9
	SS	5.5	12		SS	3.7	9
	ST	6.6	12		ST	3.9	9
Q2	US	6.7	12	Q2	US	2.4	11
	SS	7.2	12		SS	3.3	11
	ST	7.0	12		ST	3.6	11
Q3	US	4.4	12	Q3	US	5.1	9
	SS	4.8	12		SS	5.0	9
	ST	4.1	12		ST	4.9	8
Total	US	14.3		Q4	US	3.9	6
	SS	17.5			SS	2.7	6
	ST	17.7			ST	3.4	6
				Total	US	14.5	
					SS	14.6	
					ST	15.8	

The pattern of results is not so clear-cut as in the case of GCSE, but it is still generally the case that the US question got a lower mean mark than the SS and ST. There is not such a large difference between the versions as there was in the GCSE. Overall, the two papers contained about 70 marks worth - the equivalent of less than a single component (paper) at A Level. The aggregate of the structured versions is about 4.5 marks higher than the aggregate of the unstructured versions. This could be enough to have a significant effect on a full-scale exam. It can also be seen that the questions were, on average, very difficult for these candidates - the mean mark was usually at or below half marks for the question. Given that the predicted grades show that around a third of them are expected to get an 'A', this suggests that the candidates were either not well prepared or not well motivated for these trial papers.

The manipulations of structure were neither as drastic nor as wide-ranging as in the GCSE, particularly with regard to response format - there were no cases of structuring by making a question multiple-choice, or a 'cloze' style of completing the blanks. This is presumably because the nature of the subject matter made these response forms less appropriate at A Level.

There was also no real variation in the linguistic requirements among the chunks, either in the level of language to be understood in the questions, or in the level expected to be produced in the response.

There were no new forms of structuring used compared with the GCSE. The 'chunks' were larger (worth more marks) than in the GCSE so it was never the case that a single factor could be identified as varying across the levels of structure. The main differences between different versions were due to combinations of 'pure' structuring, amount of information supplied, and specificity.

1. 'Pure' structure. This varied on 9 of the 13 chunks, but it was never the sole manipulation, so we can not say what effect it had. It seems safe to assume that it had less effect than other forms of structuring.
2. Information. This varied across the levels of structure in 6 chunks, and was usually associated with better performance where more information was provided. An exception was P2 Question 2A where the US version was 'Draw a structure of the product formed when dopamine reacts with warm dilute nitric acid.' The ST version was 'When dopamine reacts with warm, dilute nitric acid, two types of reaction occur: on the aromatic ring and on the side chain. Draw the structure of the product obtained as a result of both reactions.' Performance was not significantly better on the ST version which gave extra relevant information. Performance on the whole question was very poor (see the table above), so this was probably a 'floor effect'. We might expect that with a group of more able or better prepared candidates that the structuring would have an effect. It is often implicitly assumed that the effect of structuring is to make questions easier for lower ability candidates. The results found here suggest that the effect of structuring can 'kick in' at different points on the ability scale in different questions.

3. Specificity. This varied on 6 chunks, and greater specificity was usually associated with better performance. An example was P2 Question 1B where the US version was 'Write equations for the reactions, if any, of aluminium oxide and an oxide of sulphur with an aqueous alkali and an aqueous acid'. The SS version specified aqueous sodium hydroxide and aqueous hydrochloric acid and was significantly easier. The ST version also specified sulphur dioxide rather than 'an oxide of sulphur', but this did not make any difference - performance was in fact worse (though not significantly) than on the SS version.

4. Strategy. There were 2 instances where the differences in structure gave clues to the strategy. The first was a calculation in P1 Question 2B. The US version was 'Assuming that 90% of the flue gas sulphur dioxide can be converted to sulphuric acid by the Contact process, calculate the maximum mass of sulphuric acid that could be obtained from the burning of 1 million tonnes of coal containing 1.5% by mass of sulphur.' The SS version asked explicitly for the mass of sulphur dioxide that would be produced (an intermediate step), and the ST also asked explicitly for the mass of sulphur in the coal (the first step). Performance on the US and SS versions was identical, but the ST version was significantly easier, implying that it was knowing where to start that had hindered some candidates.

   The other case where the structuring altered the strategy was in P2 Question 3A where in the US version candidates had to construct a Born Haber cycle diagram for themselves when it was given in the other two versions. The US version was significantly more difficult, although a comment from the interviews (see next section) was that having to draw the diagram helped to understand the question.

4.2.2 *Examiner's ratings of demands*

For each of the 15 question chunks in each version (the question papers and a table of chunks are in appendix) we collected examiners' rating of the demands shown in students' responses to the questions on the dimensions of Complexity, Resources, Abstraction and Strategy (CRAS). For a definition of these terms as they were used in A Level Chemistry, see the table below. The scale was conceptually a 5-point one, with points 2 and 4 defined by descriptions.

Unlike the GCSE, we did not obtain two sets of rating for demands - i.e. one set for the demands required by the question and one set for the demands evidenced in the response, but just the latter. However, the two are difficult to separate in practice and there is every reason to expect that, as in the GCSE, there would have been little difference between the two sets.

Type of Demand	Lower demand - level 2	Higher demand - level 4
Complexity	response demonstrates just one or two direct recall statements	response demonstrates linking several concepts / ideas
Resources	response demonstrates only direct use of information given in the question	candidate manipulates data or uses additional data
Abstraction	response demonstrates use of essentially factual detail	response uses explanations or calculation
Strategy	candidate follows a given series of steps	candidate selects a route to present the response

There were 180 ratings made altogether (15x4x3). As with the GCSE, the discussion below is based on comparisons amongst subsets of these ratings.

Like the GCSE, nearly all the ratings were '2', '3' or '4'. There were no ratings of '1', but two ratings of '5' were given on the dimension of Resources to US questions.

The data showing how the scale was used for each level of structure is summarised in the table below. For each level of structure, there were 15 question chunks, each rated on 4 dimensions in terms of level of demands shown in the response. This means there were 15x4 = 60 ratings made for each level of structure.

	Rating			
Version	2	3	4	5
Structured	49	2	9	0
Semi-structured	47	4	9	0
Unstructured	30	8	20	2

It can be seen that in terms of ratings, as well as mean mark, the ST and SS versions were similar. Again the majority of ratings were at the lower end of the scale, but the US versions had a significant number of higher ratings.

4.2.2.1 Differences between the different dimensions on the Scale of Demands

The number of times each dimension was rated at each score category on the scale is shown in the next table:

	Rating			
Dimension	2	3	4	5
Complexity	31	3	11	0
Resources	34	3	6	2
Abstraction	30	6	9	0
Strategy	31	2	12	0

There was no variation at all amongst the different levels of structure in terms of the rating of Abstraction, implying that (as with the GCSE) this dimension of demand is topic-related and not structure-related. The biggest differences were found in Complexity and Strategy, where the US version often got a rating of '4' compared to the ST and SS which got '2'.

4.2.2.2 Demands found using Repertory Grid technique

This was performed by one of the examiners for the GCSE Chemistry paper. The process of comparing triads of questions elicited the following constructs:

	More demanding construct		Less demanding construct
A	Quantitative	⊠	Qualitative
B	Unfamiliar context	⊠	Familiar context (i.e. book work)
C	Application to familiar situation	⊠	Application to unfamiliar situation
D	Recall / selection of information	⊠	Information supplied
E	Subject specific skills required	⊠	General skills required
F	Organisation required	⊠	No organisation required

This set of constructs seems a very plausible way of explicating the concept of 'demand'. The examiner then rated all 39 question 'chunks' from the Chemistry A Level trial papers (13 from each version) on a 3-point scale with respect to these constructs. A principal components analysis of these ratings, rotated to maximise construct loadings on the factors, yielded the following result:

Rotated solution (only loadings greater than +0.3 or less than -0.3 shown)

	Factor 1	Factor 2	Factor 3
Construct			
A		-0.72	0.43
B	0.95		
C	0.88		0.33
D		0.93	
F			0.97

Construct E, (general or subject specific skills) was excluded because 38 out of 39 chunks had received the same rating on it. Fortunately (or fortuitously) the 3 higher order orthogonal factors extracted have a relatively simple interpretation: 'familiarity' (constructs B and C), 'knowledge' (qualitative - the low demand end of construct A, plus recall - construct D), and 'strategy' (organisation - construct F, plus quantitative - construct A) were perceived to be the unrelated factors of demand applicable to these questions.

Analyses of variance of the ratings (see appendix 2 for ANOVA tables) showed that while question chunk (i.e. topic) had a significant effect at the 1% level on rating for every construct, level of structure only had a significant effect at the 1% level on constructs D and F - recall and organisation.

Of course, due to the small size of the data set, and the fact that the data only apply to the constructs and ratings of a single examiner, these findings are only suggestive - but they imply that the demands of questions that are affected by structure are the amount of recall / selection of information required, and the amount of organisation required. This is not a surprising finding, but it does help clarify how structure can affect examination questions.

Relating these repertory grid findings to the analysis of the features of structure that were manipulated we can see that the construct of recall obviously relates to the feature we called

'information'. The construct of 'organisation' can not be so easily equated to the feature 'strategy' because only a couple of chunks were seen to vary on strategy as defined in section 1, whereas there was more variation in the ratings of the question chunks on the construct of organisation. It is likely that the features we identified as 'specificity' and 'language' have contributed to the experts' ratings of organisation too. A more general question requires candidates to decide which knowledge is appropriate and to organise that knowledge into a sentence or paragraph.

4.2.3 *Students' perceptions of demands*

Two pairs of candidates were interviewed using the procedure described in chapter 2. Their views could be summarised as follows:

- Structured questions were less intimidating, because they were not faced with a large number of blank lines to fill with their own answer. It was an advantage to have the total number of marks broken down more, so they knew how much detail to put in their answer. It also made it easier to omit part of an answer if they did not know it. They did not feel that structuring necessarily made the question any easier, although it depended on the question. The general view was that in Chemistry the possible answers are so limited, in the sense that you either know the right answer or don't, that structuring makes little difference.
- Structured questions took longer to read, which may be a factor if there were question choice available.
- They felt that unstructured questions allowed more freedom to excel: "If you have extra knowledge you can put it in." The US questions also required them to monitor what they were doing.
- One interviewee felt that while sometimes the extra demands of US questions in terms of extracting and organising information may be what the question was testing, in most cases it was an unnecessary complication, testing a non-Chemistry skill. All the candidates we interviewed were also doing an essay-based subject for A Level, but they all said they were not concerned about their use of English or powers of expression when doing Chemistry, and felt they could get away with simpler note-like sentences.

Overall, their preference was for the semi-structured or unstructured versions, which they felt were more appropriate for A Level, although they did welcome the greater breakdown of marks given on the structured versions. It should be borne in mind that the interviewees were all high ability candidates and their views may not be entirely representative.

4.2.4 *The effects of structure on demands in A Level Chemistry*

1. In terms of the effect of structure on performance, the same effects were found as at GCSE, only to a much smaller extent, so it is less valid to draw any confident

conclusions. In general, reducing the structure did make questions more difficult - i.e. lowered the mean mark.

2. No new aspects of structuring were found, and there was almost no variation in language or response format. Most differences were a combination of pure structure, information provided, and specificity.
   Providing extra information and increasing specificity usually, but not always, improved performance.

3. Using the modified Edwards scale, most ratings of questions on all versions were at the lower end of the scale of demands. The ST and SS versions were usually given the same rating. Most of the higher ratings were given to the US version.

4. As with the GCSE, the dimensions of demand that varied across levels of structure were Complexity and Strategy. Abstraction was not related to level of structure.

5. The Repertory Grid analysis identified 6 'constructs' relating to demands, which a principal components analysis reduced to 3 uncorrelated factors, which could be described as 'familiarity', 'knowledge', and 'strategy'.

6. The candidates found structure helpful in giving a greater breakdown of mark allocation, but sometimes rather restrictive. They felt that in general it did not have much effect in A Level Chemistry where knowledge of the material is the decisive factor.

Inspection of the question papers used in the trials shows that the differences between the versions at A Level were much less radical than at GCSE, so it is not surprising that much smaller differences in performance and rating were found. With hindsight, it is possible that the constraint to make the mark schemes for each version as similar as possible was too tight at A Level. This may be because the knowledge and skills being tested are more specific and it is harder to get at them in different ways within the same question.

4.3 Combined summary for GCSE and A Level Chemistry

1. Making a question less structured makes it more difficult, in terms of lowering the mean mark. At GCSE the effect was noticeable even at an individual question level, and accumulated over a whole paper could make a difference of at least 12 marks out of 100. Increasing the amount of structure could increase the mean mark by the same amount. At A Level, the same effects were found as at GCSE, only to a much smaller extent.

2. At GCSE, six aspects of structuring could be identified, five of which affected performance. However, the 'purest' form of structuring - breaking the question into several parts without providing or requesting any extra information - had no significant effect on performance. At A Level there was less variety in types of structuring used. Most involved a combination of 'pure' structure, information provided, and specificity. Providing extra information and increasing specificity usually, but not always, improved performance. In general, with the exception of 'pure' structuring, changing the structure

affects the nature of the question. The different types of structuring overlap, so it is hardly ever possible to vary a single aspect.

3. In GCSE Chemistry, there was no significant difference between the demands required by the question and the demands candidates were able to show in their response. The implication is that to test higher-order skills such as linking ideas and planning a strategy, less structured questions need to be set.

4. Using the modified version of Edwards' scale of cognitive demands, most ratings given to all versions of all questions were at the low demand end of the scale. Most of the higher ratings were given to unstructured questions. Manipulations of structure altered the demands on the dimensions of Complexity and Strategy, but not Abstraction or Resources.

5. Complexity relates mainly to the need to make links between ideas and to demonstrate understanding rather than straight recall. Strategy relates mainly to the requirement to devise and monitor a strategy. The level of Abstraction did vary between questions, but not across the levels of structure, implying that this 'demand' is topic related rather than structure related.

6. At A Level, the repertory grid analysis identified six 'constructs' relating to demands, which a principal components analysis reduced to three uncorrelated factors, which could be described as 'familiarity', 'knowledge' and 'strategy'.

7. From the candidates' point of view structured questions are less intimidating and generally easier. They like to have a greater breakdown of the mark allocation, and like it to be made clear what the expected response to a question is. They recognised that some questions were more appropriate in an unstructured form, and that for a candidate on top of the material it might be easier to gain marks on this type of question. A Level candidates felt that structure did not really have that much effect since Chemistry is a subject where knowledge of the material is the decisive factor.

5. HISTORY

5.1 History GCSE

MEG Modern History GCSE (syllabus 1607) was chosen as a basis for the trial questions.

Note: *Candidate:* a person taking an examination *Student*: a person studying for an examination

5.1.1 *Structuring the questions*

The task of the examiner was to write three versions of History GCSE questions which were as similar as possible in subject content, and varied only in terms of the structure of the question.

The examiner expressed concern at writing unstructured questions for GCSE students. These students were more used to structured questions and he felt that they would be unable to cope with open ended questions which did not guide them through the process of answering. He predicted that students attempting the unstructured questions would fail to include the relevant material (which, in a more structured question would be suggested by prompts). He suggested that candidates attempting the unstructured questions would finish answering after 10 minutes (of a 50 minute exam) because they were not either (i) prompted to include certain historical ideas/content and (ii) led through the process of answering the question.

Despite his reservations he was persuaded to create three versions of a question which aimed to cover the same historical material. The questions are shown in appendix 3.

Structured questions varied from the other versions by the degree of support the candidate was given in reaching the objective of the question, which for example in Question 1 was to evaluate and analyse a given historical source. The structured version led the candidate through lower level skills of description and factual recall before presenting the final sub-question which required evaluation. The unstructured questions characteristically asked candidates to evaluate without having guided them through the process of considering the information necessary to make that evaluation. For example, the unstructured version of Question 2 asked candidates 'How successful was the League of Nations in the period 1920-1939? Explain your answer.' The structured version of this question led candidates through four sub-questions with the ultimate aim of considering how a number of given factors contributed to the failure of the League of Nations.

5.1.2 *The effects of structure on difficulty*

This time our procedure was not as successful at achieving matched ability groups across versions as in the other subjects. Mean predicted grades for the students who answered each question are reported below; with Grade A equal to 8 and Grade U equal to 1.

	Structured	Semi-structured	Unstructured
Question 1	6.51	5.48	6.66
Question 2	6.66	6.51	5.48
Question 3	6.62	6.49	6.30

The table shows that on Question 3 the three versions were taken by students of about the same ability. However, the students taking Question 1 (Semi-Structured) and Question 2 (Unstructured) were predicted to score about a grade less than the other groups, and we must accept that ability differences may influence the results of the analyses, with candidates sitting Question 1 (Semi-Structured) and Question 2 (Unstructured) expected to perform worse than the other students.

5.1.2.1 Mean marks

The mean mark for each version of each question is given below.

		Unstructured	Semi-structured	Structured
Q1 (max 15)	mean	9.05	9.96	9.87
	sd	2.43	2.72	2.33
Q2 (max 20)	mean	10.10	8.94	11.45
	sd	5.13	4.63	3.60
Q3 (max 30)	mean	17.73	14.60	16.72
	sd	5.70	5.31	4.96

Analysis of variance showed that in Questions 3 and 2 structure had a significant effect on performance ($F_{(2, 174)} = 5.363$, $p<0.005$) and ($F_{(2, 173)} = 4.782$, $p<0.01$) while Question 1 showed no effect ($F_{(2, 173)} = 2.285$, $p<0.105$).

5.1.2.2 Question 1

There was no effect on structure in Question 1. Students performed equally in all 3 versions.

5.1.2.3 Question 2

Students performed similarly on the unstructured and semi-structured versions, and better on the structured version than either of the other versions (average marks close to 60%). This begs the question 'how are the unstructured and semi-structured different to the structured version?' Explanations may be found in the following features of the questions and mark schemes:

- *Structuring the process*: In the semi-structured question the process of answering the question was broken down, into the two decades (the 20's and 30's), and given this structure the students had more scope for devising their own strategies than those attempting the structured. In the structured question the strategy for answering the question was largely decided for the students so they were less likely to devise or use inappropriate strategies.
- *Giving prompts to relevant material:* The US and SS questions did not give the students clues to the material which could be included in an answer. The structured version, on

the other hand, provided specific examples for the students to write about. This would help students by removing the need to identify or recall relevant material.

- *The unstructured question was dense*: In the US question students were expected to complete several processes at the same time, whereas the S version involved simple operations and ideas (albeit more of them).
- *The unstructured question required evaluation and higher level skills:* The US and SS versions asked candidates 'How successful' they considered the League of Nations to be, whereas the S version told students that the League was 'a success in the 1920's' (part c) and that it was a 'failure in the 1930's' (part d). Thus the students were required to evaluate the material in the US and SS versions.
- *Gaining marks for recall:* The earlier parts of the structured question (parts a and b) tested little more than recall and deployment of information, which would allow students to gather a number of easier marks.

5.1.2.4 Question 3

Students attempting the structured and unstructured questions performed significantly better than those sitting the semi-structured. This was a surprising result with candidates sitting the US question performing best. An interrogation of the nature of the question and mark scheme suggested reasons for this finding:

- *High ability sample*: The sample of students sitting this question was of slightly higher ability (almost a third of a grade higher) than those sitting the other two questions (although there was a lot of overlap in the samples). The higher ability students were able to cope well with the unstructured question.
- *Opportunities for higher level skills:* The structured question did not enable students to make connections, synthesise or reach overall judgements, unlike the other two questions.
- *Using their own ideas:* There was more opportunity in the US and SS versions for the student to bring in ideas than in the S version.
- *A change in approach of the sub-questions:* The structured and semi-structured versions supported students throughout the first question parts, building up to a final sub-question which gave them more freedom. These final questions required that candidates changed their approach from one of being supported and responding to specific demands to an approach in which they were required to synthesise and evaluate making their own judgements. When students were required to switch from working with support to working without it, they had difficulty with the question. This resulted in a better performance on the unstructured version than the others.
- *Unstructured question gave students freedom:* Students were given freedom to use a strategy as they saw appropriate. The most able of them could devise and implement a

strategy which suited their knowledge and the requirements of the question. Also if they could not recall the required information they were not able to introduce other material as a substitute to gain marks.

- *The structured question was very specific:* In the structured question there was not much scope for candidates to show that they could decide what material was relevant and irrelevant.

5.1.2.5 The effects of structure on demands on GCSE History

Three features of questions and mark schemes appeared to affect performance:

1. **Prompts to relevant content may** reduce the demand on students to decide which material was relevant and thus helped candidates.
2. Questions which give students **scope to devise their own strategy** can be more demanding than those which provided a strategy.
3. Questions which provide more **opportunities for higher level skills** (for example, evaluation, analysis, complex explanation and making generalisations) can be more demanding than questions which give marks for lower level skills (for example, recall).

A final issue, although not a feature of the question and mark scheme, but nonetheless a significant influence on performance on the different versions was the ability of the students. The higher ability students were better able to take up the opportunities given by questions. The higher ability students were (i) less reliant on prompts to relevant content; (ii) better able to take up the challenge of devising an effective strategy and; (iii) better equipped to take up opportunities to show their higher level skills.

These features were further investigated using data from (i) examiners' descriptions of demands in questions and ratings of responses and (ii) interviews with students. The next two sections present that evidence.

5.1.3 *Examiner's rating of demands*

The examiner applied the CRAS scale of demands to the questions (see page 6). The task was to rate (i) the questions and (ii) the candidates' responses on the four dimensions - complexity, resources, abstraction and strategy.

5.1.3.1 Rating questions

The examiner found it difficult to rate History questions in terms of demands because, he explained, History questions do not make *demands* on candidates, they *enable* candidates to reach a level of performance by giving *opportunities*; there is therefore not a ceiling of demand in a History question which could be given a rating. When considering the four dimensions outlined by the scale it is possible that a History question could help a candidate to perform at

least at the highest level. In fact, it is really the mark scheme which makes demands, as it defines the level of performance required to gain marks. The examiner commented on how the opportunities were different across the versions. To the candidate, of course, there is no difference between demands set by the question and those set by the mark scheme. The demands, or opportunities, described by the examiner are presented in the next table.

	Complexity	Resources	Abstraction	Strategy
Q1	S version required the recall of specific knowledge.		There was more scope for dealing with ideas in US than in S.	Little opportunity for candidates to develop strategy in S. S gave limited scope for them to organise their information. SS much more scope and US given free rein.
Q2	S and SS version did not require that candidates brought all things together and concluded at the end. S and SS led candidates more gently into the task	S provided specific examples for students to write about. US gave no information, but gave freedom.	Opportunities were more limited in S version. Opportunities were about the same for SS and US.	Scope for devising strategy increased through S, SS and US. S version set limits beyond which it was difficult for student to go.
Q3	Students only required to carry out one thing at a time in S version. So S didn't give opportunity to make connections.	Little scope for using resources in any way other than that directed in S question.		Structure was given to candidates in first parts of S version. Fewer demands in S version. Higher demands in US version.

In 10 out of the 12 cells dimensions varied across the three versions. The questions were seen as different in their demands on all four dimensions.

Structured questions were usually seen as less demanding than semi-structured or unstructured versions. Of the 10 cells that were different across the versions, 8 of them showed that the structured version was likely to make fewer demands. There were two exceptions to this, in those cases the structured questions were predicted to be more demanding. In these instances, the responses were poorer from the candidates doing the structured versions because they had been restricted to the prompts that the question had given them (for example Question 1 complexity and Question 3 resources). It seems that the candidates found themselves restricted to the specific requests of the question and were not given the chance to show their capabilities beyond the remit of the question. This is also reflected in the comments of the students we interviewed.

The complexity of a task, the use of resources and the level of abstraction were usually the same (or very similar) across the structured, semi-structured and unstructured versions). Where there were differences between the versions they tended to be in the strategy of the questions, i.e. the demands made on candidates to organise their own argument and answer.

How does the data on the opportunities in questions relate to the three features of the questions which affected performance?

1. *Prompts to relevant content.* The examiner predicted that the structured versions of Question 1 and Question 2 would make different demands compared to the other versions because they required the recall of specific knowledge. He was unable to predict whether this would aid or inhibit performance.
2. *Scope to devise their own strategy.* In all three questions the unstructured questions were described as giving more opportunities for students to develop their own strategies than the structured or semi-structured questions.
3. *Opportunities for higher level skills.* There were more opportunities for candidates to demonstrate their higher level skills in the unstructured questions. The US versions of Question 1 and Question 2 allowed candidates to include abstract ideas. The US versions of Question 2 and Question 3 encouraged synthesis and linking.

5.1.3.2 Rating responses

After marking the responses the examiner rated the answers on the four CRAS dimensions. The examiner also wrote a report on the responses. The table below shows that the questions were rated as making different demands in terms of all four dimensions: complexity, resources, abstraction and strategy. The text in the table is drawn from the examiners report on the responses.

Question	Complexity	Resources	Abstraction	Strategy
Q1 Rating	S SS US 3 3 2	S SS US 3 3 2	S SS US 2 4 2	S SS US 3 5 3
	Few US students were able to both investigate and evaluate the source as well as analyse reasons outside of the source. Specific material in the structured question made it very difficult.	The resource was not used as effectively in the US question as the S question. Students attempting the US version did not investigate the source in the same detail as those who did the S question.	SS version was more likely to bring in abstract ideas than either the S or the US versions.	SS responses showed candidates were better able to devise, monitor and organise their strategy for answering. US candidates struggled with having no clear strategy at all. The SS gave more freedom, which led to a variety of effective strategies.

Continued

Q2 Rating	S 3	SS 1	US 3	S 4	SS 1	US 2	S 4	SS 2	US 2	S 4	SS 3	US 2
	SS showed lack of explanation and generalisation and tended to focus on description.			SS version made poorer use of data and information (given or recalled). SS question broke down the strategy for answering rather than helping with the content. SS students had to generate much of the information themselves.			More responses to the structured question attempted to include abstract ideas.			The open question gave more opportunities for candidates to devise and monitor their own *strategies*. Only the very able candidates could do this effectively.		
Q3	S	SS	US	S	SS	US	S	SS	US	S	SS	US
Rating	3	2	3	2	4	4	4	4	4	2	4	4
	The evaluation in the US responses was superior to SS.			US were faced with all the sources at once, with an open question requiring the evaluation of all sources. This appears to have given them opportunities rather than presented difficulties.			Abstraction was not affected by the level of structure.			US version gave some very good results with students organising a large amount of material effectively using a surprisingly large number of sophisticated and effective strategies.		

Question 1

The structured version of this question made specific demands on students, requiring that they identified the five figures in the cartoon (part a) as well as dealing with two particular aspects of the Treaty of Versailles (part c). In part a, only recall was required, and a few students could not recall that necessary information which, for those students, made the question impossible. However, this was a small minority of the students and most responded well to the specific question and scored high marks.

The higher level skills that this question was aiming to draw from students were explanation (within the historical context) and evaluation of the source. There were also opportunities for dealing with abstract ideas. The ratings showed that students attempting the semi-structured version were more likely to employ these higher level skills. There was some evidence of them in the unstructured responses, but candidates tended to use only <u>one</u> of the higher level skills, rather than both. This was because they were required to complete several processes at a time and one became lost in the process.

The semi-structured responses showed better organisation of material compared to the structured responses which although organised at a basic level did not effectively link ideas. The unstructured responses showed very little organisation and the students obviously struggled.

Overall, the rating of the responses to Question 1 reflected the statistics which showed that candidates attempting the semi-structured version performed best, if only by a small margin.

Question 2

Mirroring the statistical findings, the examiner's ratings consistently showed that the structured version elicited the best performance. The structured question gave more prompts to relevant content, whereas the semi- and unstructured versions gave no prompts to content.

Although the unstructured question gave more opportunities for the development of strategies, only the very able candidates were able to do so. The answers to the structured questions were well organised and ideas were linked, although it was difficult for students to go beyond the limits set by the question.

There was scope for higher level skills in all three version of the question (even in the structured question, although only in its latter parts). However, because the semi-structured answers lacked effective selection and organisation, few students were able to produce explanation at a high level. The best students attempting the structured and unstructured versions did include complex explanation and judgement in their responses.

Overall in Question 2 the provision of prompts to relevant content and the support given in presenting an answer helped students generate good answers fulfilling the higher order skills required by that question.

Question 3

The examiner's comments about the opportunities given by the questions suggested that the structured question made fewer demands on students because it gave fewer opportunities for making links between ideas, using the resources in any way other than that directed by the question, and devising a novel strategy. Despite this, it was the unstructured version which gleaned the best responses. The examiner reported that the performance on the unstructured version was, at the top range of ability, better than that on the other versions.

The very able candidates attempting the unstructured question were better able than those restricted by the structured or semi-structured versions to show their abilities to select and recall relevant material. The structured question directed students to the relevant sources and they tended to keep to the sources to which they had been directed rather than cross-referencing to other sources. Students attempting the unstructured question, on the other hand, used most or all of the sources and were able to discriminate relevant and irrelevant material.

Despite not having the support given to the other students, those sitting the unstructured version were able to analyse and evaluate sources and present their answer using an effective strategy.

5.1.3.3 Summary of examiner's ratings of demands

How did the examiner's ratings of the responses relate to the three types of structure which influenced performance?

1. *Prompts to relevant content.* Responses to the unstructured version of Question 1 were less detailed than responses to the structured version. The structured version gave some specific prompts to candidates about what material to include in their answers. The effect of the prompts to was to (i) prompt recall and (ii) provide relevant examples.
2. *Scope to devise their own strategy.* In Question 1 responses to the semi-structured question showed better devising and monitoring of strategies. In Question 3 the best strategies came from students taking the unstructured question. This was not as the examiner predicted. Although those attempting the US question were not given any strategy to use they were most effective in using the freedom given to them to devise a variety of sophisticated and appropriate strategies.
3. *Opportunities for higher level skills.* The examiner's analysis of the questions suggested that there were more opportunities to show the higher level skills in the unstructured questions. The students' work confirmed this: Evaluation in the unstructured versions of Question 2 and Question 3 was superior to that in the semi- and structured versions. However, responses to the semi-structured and structured version of Question 1 and Question 2 were more likely to bring in abstract ideas.

The three features of questions had different effects on performance depending upon the ability of the students. The sample taking Question 3 were higher ability than those taking Question 1 or Question 2. The unstructured version of Question 3 gave more opportunities to candidates and they were equipped to take up those opportunities. This is supported by the statistical analysis, and interviews with students, as well as the examiner's perceptions discussed above.

5.1.3.4 Demands found using the Repertory Grid technique

The repertory gird technique allowed the examiner to describe what he thought were the demands in the questions. Two types of information were gleaned from the repertory grid data: Firstly, constructs were elicited from the examiner. These were the examiner's description, in his own language, of the characteristics of the questions that affected their difficulty. Secondly, some questions were rated on a five point scale on each of these constructs. The data from the examiner's ratings explained the differences and/or similarities between the questions, and between the constructs.

	More demanding	Less demanding
1	The candidate gets straight in.	Candidates are led through the question.
2	Reach overall judgement.	Never quite asked to pull the whole thing together.
3	Asked for specific knowledge.	Not necessarily asking for specific knowledge.
4	Having to select relevant information for themselves.	Helpers are a trigger to recall and select relevant knowledge.
5	Abstract	Asked to do concrete/surface things. They can accumulate marks for lower level skills.
6	Candidates have to perform pedestrian/ mundane things throughout the question. They need perseverance.	Mark scheme is broader, looking for quality of answer.
7	Given the opportunity to look at both sides.	Not given that opportunity.
8	Decide criteria for judgement.	Candidate doesn't have to decide on criteria for judgement. It is provided.
9	Expected to make links, but not instructed to.	The scope for making complex links is less. Or they are told to make links.
10	Develop logical arguments.	No scope for logical argument/ steps.
11	Everything is implicit. They have to mark their own decisions about their moves and strategy.	The strategy to use is made explicit and candidate is given small chunks to work on.
12	They must manage everything.	Some issues have been raised in a previous questions, so the previous question informs the final evaluation.
13	The candidate has to cross reference the sources for themselves.	Differences and similarities are pointed out.
14	All sources not individually interrogated before final evaluation.	They have written about every source before evaluating, so the resources have received a more detailed interrogation before evaluation.
15	Mark scheme is explicit.	MS can't expect all moves of the structured question. Mark scheme is not an accumulation of points.

5.1.3.4.1 *Rating questions on the constructs*

The examiner rated two of the three questions on each of the constructs. He rated Question 2 (the three versions of the question on the League of Nations) and Question 3 (the three versions of the question about the Marshall Plan in the Cold War).

A rating scale of 1-5 was used, where a rating of 1 was given for the higher level construct (the left hand column of the preceding table) and a rating of 5 for the right hand column.

Factor analysis of the ratings showed the relationships between the questions and the relationships between the constructs.

Relationships between the questions

Those that were seen as most similar in terms of the constructs were:

- Question 3 unstructured and Question 2 unstructured were factored with the negative of Question 1 structured, which means that the structured question was seen as very different to the unstructured questions. This suggests that structure was an influential factor in the examiner's descriptions of the questions.
- Question 3 structured and Question 3 semi-structured. These questions were seen as similar, probably because they were both on the same topic and neither gave clues to relevant content.

	Construct	Factor 1	Factor 2
1	straight in	.97	
2	overall judgement	.70	.63
3	specific knowledge	-.90	-.42
4	select information	1.00	
5	abstract	.73	.64
6	perseverance	-.92	
7	both sides	.73	.64
8	decide criteria	.90	
9	make links	.97	
10	develop arguments	.89	.30
11	strategy explicit	-.94	
12	previous question	-.89	.39
13	cross reference	.61	-.62
14	sources interrogated	.48	-.80
15	explicit mark scheme	-.94	

There were two factors of demands:

Factor 1: The extent to which the candidate is led through the question, being given clues about (i) how to organise and present their answer and (ii) what would be relevant material to include in their answer. (Constructs 1, 3, 4, 7, 11, 12, 13, 14 and 15.)

Factor 2: The extent to which higher order skills (e.g. evaluating, concluding and defining their own criteria for evaluation) are required. (Constructs 2, 6, 8, and 10.)

Factor 1 accounted for 72.2% of the variance and factor 2 accounted for 19.7%. This shows that factor 1 was the most important to the examiner.

5.1.3.5 Summary of Repertory Grid Interview results

The two factors seen as most important to the examiner overlap with the three features of questions previously identified as affecting their demands. Factor 1 is an accumulation of the two features (i) specificity of content and (ii) specificity of strategy. Factor 2 relates to (iii) the opportunities in a question for higher level skills.

5.1.4 *Students' perceptions of demands*

Interviews with students confirmed that they also see as important the three features of questions identified in the statistical analyses and examiner's ratings.

5.1.4.1 Prompts to relevant content

Students saw specificity as a double edged sword: They appreciated the support a specific question gave them, as one student said 'you can pick out the relevant information that's wanted'. Yet there was an awareness that a question which prompted specific information could be limiting because, as one student commented 'you can't put your interpretation on it'. Higher ability students tended to be concerned with the latter of these two points. They would have preferred less specificity because they were 'less restricted' and a specific question could encourage you to 'put too much detail in and use up all your time'. For example, a question asking about two particular historical disputes could cause difficulties: 'if you haven't done these disputes then just having these could be confining - either you know it or you don't'. Despite the high ability students preferring the unstructured questions they still described the structured version as easier. Lower ability students, on the other hand liked a structured question because it 'gives you pointers to things you would forget'. Structured questions with prompts to material that should be considered in the answer were preferred because they 'squeeze out all the little details and jogs your mind' and were 'more broken down' allowing the students to 'concentrate on one thing at a time'.

Unstructured questions were described as more difficult because 'you use your own knowledge'. Structured questions, on the other hand, were thought to be easier because 'it's easier to do more questions with fewer points'. The implication being that the more small questions there were to attempt the more chance there was of gaining marks. They also noted that with an open question there was 'only one chance to get the question right' and if you didn't understand the question then you were completely stuck. Structured questions were preferred because they had prompts (or 'helpers' as they called them) to the content that should be included in their responses.

5.1.4.2 Scope to devise their own strategy

They would use the mark allocation to decide where to put the detail, and where to expand upon points. So, as the structured version has a more detailed mark allocation this meant that the structured questions were more helpful in allowing them to allocate time to each sub-question. As one student said 'it helps you pace yourself and time your answers, but with the green (unstructured) you are likely to go over time and write too much'.

Students preferred the structured question which gave them a structure to use, and one commented that when answering the structured version 'you stick to what you are asked and so you are less likely to go off track'. The structured version was described as having 'steps' which 'structure it for you'. This was preferred because it allows you to 'get into the examiner's mind to see what he or she expects'.

5.1.4.3 Opportunities for higher level skills

The more able candidates interviewed preferred questions which allowed them to show their knowledge and skills. For example, one very able candidate described how the structured version 'wouldn't let me say as much as I wanted, especially about the effect of the League on WWII and what could have been done to make the league more successful'.

5.1.4.4 Summary of interview findings

Students preferred to have prompts to relevant material, although they were aware that the more specific the content had to be the more important it was that they could recall the related knowledge. Most students found questions which required them to devise their own strategy for answering very demanding, however the most able students preferred more open questions which allowed them a free rein to include what they wanted and organise it as they saw fit.

5.1.5 *The effects of structure on demands in GCSE History*

It was not universally the case that structure of questions increased performance in History GCSE. However, three issues consistently had an effect on performance.

1. The provision of prompts to relevant material.
2. The provision of scope for students to develop a strategy for responding.
3. The opportunity a question gave to use high level skills (such as making generalisations, synthesising, evaluating and analysing).

What effect these features had was mediated by the ability of the student doing the question.

- **The provision of prompts to relevant material.**

For the lowest ability students very specific questions giving prompts to relevant material could act as a barrier. If they could not recall the particular information they couldn't gain any marks. For average ability students the prompts acted as an aid by prompting recall and pointing to relevant material to include. For the most able students (those predicted to gain grades A and B) specific questions restricted them and prevented them from showing that they could make decisions about what content was relevant or irrelevant.

- **The provision of scope for students to develop a strategy for responding.**

The highest ability candidates thrived on questions which gave them scope to develop their own strategy for organising and presenting their answer. The average ability candidates coped best with a question which gave them a strategy to follow.

- **The opportunity a question gave to use high level skills (such as making generalisations, synthesising, evaluating and analysing).**

Average ability students benefited from some support and guidance to build up from low level skills (for example recall) to higher level skills of evaluation analysis and making

generalisations. On the other hand the best students would take the opportunities given to them by unstructured questions to demonstrate their higher level skills.

5.2 History A Level

History A Level syllabus 9020 includes 18 papers of which students are entered for two (with various restrictions). The most popular options are English History 1450-1714 and European History 1450-1715. These options were used as a basis for this study.

5.2.1 *Structuring the questions*

Essay questions are a common means of assessing History at Advanced level. One objective of the syllabus is to test candidates' proficiency in '...the ability to present a clear, logical, concise and relevant argument.' (OCEAC 1996). Essay questions are used as a means of assessing these skills.

There were two ways in which the questions were structured by the examiner: (i) by adding supports to guide the student through the *process* of answering and (ii) giving prompts to the *content* of the answer. The former of these was used more frequently to structure the A Level History questions. The questions trialled can be found in appendix 3.

5.2.2 *The effects of structure on difficulty*

Three versions of eight History questions were trialled in 37 schools in England. There was a choice of questions, candidates answered two out of eight questions (one from questions 1-4 and one from questions 5-8).

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
n	190	35	89	35	41	65	181	47

The most popular questions were Question 1 (Henry VII), Question 3 (Mary I) and Question 7 (Ferdinand and Isabella).

The next table shows that forecast grades did not vary significantly across the three versions in each set of questions.

	Structured	Semi-structured	Unstructured
Qs 1-4	5.31	5.47	5.23
Qs 5-8	5.51	5.26	5.42

The average predicted grade for the students sitting the versions was similar (low to middle C).

5.2.2.1 Mean marks

The mean marks scored in each version of each question were as shown below.

		Unstructured	Semi-structured	Structured
Q1	mean	12.91	13.67	12.98
	sd	3.18	3.38	3.28
Q2	mean	14.25	13.0	14.00
	sd	3.93	2.91	3.87
Q3	mean	14.17	12.96	14.28
	sd	3.11	2.34	3.74
Q4	mean	14.00	12.58	12.36
	sd	2.21	3.07	2.29
Q5	mean	12.08	11.88	13.00
	sd	3.75	4.41	3.85
Q6	mean	14.85	12.58	14.58
	sd	3.98	3.32	4.13
Q7	mean	14.43	12.25	12.58
	sd	2.97	3.19	3.63
Q8	mean	13.90	14.33	13.75
	sd	2.77	2.43	2.21

Analysis of variance showed that there was no significant effect of structure on performance for (questions 1-4: $F_{(2,3)}$=0.991, $p<0.37$; questions 5-8: $F_{(2,3)}$ =1.037, $p<0.38$) demonstrating that the level of structure had no effect on level of performance. For questions 1-4 there was no significant effect of choice of question on performance (showing that candidates performed equally well across questions 1, 2, 3 and 4). However, there was a significant effect of choice of question on performance for questions 5-8 ($F_{(2,3)}$ =2.779, $p<0.04$). Candidates performed best on Question 6 and worst on Question 5. This suggests that the topic of the question was more influential on performance than the level of structure of the question.

Two types of structuring were employed to create three versions of the original essay questions. These were: (i) the provision of a strategy for answering and (ii) the provision of prompts to the content of an answer.

All of the eight structured questions provided students with a strategy for answering. For example, the unstructured version of Question 4 asked

> *How far were the problems James I faced as king the result of his weakness of personality?* *[25 marks]*

The semi-structured version helped students organise their answer by breaking the question down into two parts:

> *(a) What were the main problems James I faced as king?* *[5 marks]*
> *(b) How far were these problems the result of this weaknesses of personality?* *[20 marks]*

The structured question consisted of four sub-questions:

> *(a) What were the main problems James I faced as king?* *[5 marks]*
> *(b) In what ways can it be argued that these problems resulted from his weaknesses of personality?* *[8 marks]*
> *(c) What other explanations can be advanced for these problems?* *[8 marks]*

(d) How far were these problems the result of his weaknesses of personality?
[4 marks]

The expected content of the answers was very similar, but the more structured questions gave students a strategy for presenting and organising their answer. Providing a strategy was the most common way of structuring the questions. The second means of structuring was to give prompts to relevant material to include in their answers. Only three of the eight questions used this type of structure, these were questions 1, 6 and 7. Questions 1 and 6 suggested particular material to include in their answers. For example the unstructured version of Question 1 asked:

> *What were Henry VII's aims in foreign relations and how successful was he in achieving them?* *[25 marks]*

The structured question gave more support for candidates in both organising their answers and deciding what material was relevant. Part (b) of the question gave prompts to content:

> *Explain what policies he pursued to achieve these aims. You might wish to refer in your answer to relations with France, Spain, the Netherlands and Scotland, but you need not confine yourself to these.* *[15 marks].*

Overall in A Level History, the provision of an answering strategy was a more popular way of structuring the A Level questions than providing specific pointers to relevant content.

Question 3 in the GCSE study provided organisational structure rather than content structure, this type of structuring did not help students as much as the content structure in GCSE Question 2. The evidence suggests that providing structure for the content of answers aids performance more than providing organisational structure.

Further investigation of (i) students' perceptions of the question and (ii) the examiner's ratings of demands in the questions and his judgements of the quality of responses could explain why structure should not impact on performance on History A Level questions.

5.2.3 *Examiner's rating of demands*

5.2.3.1 Scale of demands

The examiner applied the CRAS scale of demands to the questions. The scale is described in the introduction. However, the examiner adapted it for A Level History.

		Lower demands.		**Higher demands.**	
Rating	**1**	**2**	**3**	**4**	**5**
Complexity		Requires only elementary links and /or evaluation. No requirement to put two sides of an argument	↔	Requires discussion of links. Focuses on evaluation. Requires balanced arguments.	
Resources		Knowledge required primarily 'factual' - events, dates etc. Some inaccuracy and omissions permissible. Selection (relevance) may be obvious from question.	↔	Requires knowledge of concepts as well as 'facts'. Knowledge accurate and, where necessary, detailed. Requires developed sense of relevance.	
Level of Abstraction		Requires narrative or description and focus on events rather than ideas. Avoids need for technical terms.	↔	Requires analysis and explanation rather than narrative or description. Requires grasp of technical terms.	
Strategy		Provides clear and detailed framework for answer.	↔	Requires candidate to organise own argument.	

The examiner rated (i) the questions and (ii) the candidates' responses on the four dimensions of complexity, resources, abstraction and strategy.

5.2.3.2 Rating questions

The comments made in the report of GCSE History stand also for A Level History. To summarise: the rating of the questions was problematic because History questions do not make *demands* on candidates, they *enable* candidates to reach a level of performance by giving them *opportunities*. The examiner commented on the opportunities given in the questions and rated the students' responses using the CRAS scale.

The next table shows that the majority of differences in demands related to the dimension 'strategy'. In six of the eight questions the demands of the strategy dimension were predicted to be higher in the unstructured than in the structured question. The examiner thought that the provision of a structure would give candidates greater direction whereas the unstructured questions would be more likely to cause problems of devising or maintaining a strategy and therefore presenting a coherent argument: there is also an implication on the resources dimensions arising from more specific demands on knowledge that is left implicit.

Question	Complexity	Resources	Abstraction	Strategy
Q1 Henry VII				S and SS provided more supports than US version.
Q2 Somerset	S and SS were less demanding than US. Part (a) was given as a foundation on which to base argument.			S and SS less demanding because part (a) was given as a foundation.
Q3 Mary I				
Q4 James I	Some pointers to other sides of argument were given.			S would help candidates organise their argument more than US.
Q5 Suleiman			Content of US was more abstract than SS and S.	US presented greater problems in devising a strategy than S and SS.
Q6 Francis I		S gave pointers to what information to include.		Pointers given in S version aided organisation of answer.
Q7 Ferdinand and Isabella	SS did not ask for links to be made between the two main ideas.			SS version did not require that links were made.
Q8 Philip II				

(Blank cells denote that the examiner saw no differences in opportunities between the three versions of the questions.)

The complexity of the task, the use of resources and the level of abstraction were usually the same (or very similar) across the structured, semi-structured and unstructured versions. Where there were differences between the versions they tended to be in the strategy (i.e. the demands made on candidates to organise their argument and answer). This relates to previous discussion of the type of structure used in the History A Level questions. Despite this variation in the perceived demands of questions there were no significant differences in performance. We can conclude that the type of structuring used in the A Level (i.e. structuring the strategy) did not impact upon performance.

5.2.3.3 Rating responses

Once the scripts had been marked the examiner rated the level of complexity, resources, abstraction and strategy demonstrated in the responses. The ratings are shown below.

Question	Complexity	Resources	Abstraction	Strategy
Q1 Henry VII				
Structured	3	4	3	3
Semi-structured	3	4	3	3
Unstructured	3	4	3	3
Q2 Somerset				
Structured	**2**	3	3	**2**
Semi-structured	**3**	3	3	**2**
Unstructured	**3**	3	3	**3**
Q3 Mary I				
Structured	3	**3**	3	**3**
Semi-structured	3	**4**	3	**3**
Unstructured	3	**4**	3	**2**
Q4 James I				
Structured	3	3	3	**3**
Semi-structured	3	3	3	**2**
Unstructured	3	3	3	**2**
Q5 Suleiman				
Structured	**3**	3	**3**	**2**
Semi-structured	**3**	3	**3**	**2**
Unstructured	**2**	3	**2**	**1**
Q6 Francis I				
Structured	3	**3**	2	2
Semi-structured	3	**3**	2	2
Unstructured	3	**2**	2	2
Q7 Ferdinand and Isabella				
Structured	**4**	2	3	3
Semi-structured	**4**	2	3	3
Unstructured	**3**	2	3	3
Q8 Philip II				
Structured	3	4	3	**2**
Semi-structured	3	4	3	**3**
Unstructured	3	4	3	**3**

Ratings were very close and any differences between the three versions were of one point on a five point scale. It was very rare for the examiner to use the extremes of the scale. Of the differences in ratings most were on the strategy dimension. This shows that responses varied on the extent to which the students devised and maintained a strategy for tackling and answering the question. Responses, as predicted, were less likely to vary on complexity, resources and level of abstraction.

The examiner predicted that unstructured questions would require that candidates organise their own argument and monitor their answers, whereas the structured questions were more likely to support candidates in devising and maintaining a strategy. There were differences in responses on the strategy dimension, although these were not consistently in either direction: sometimes providing a framework for responding resulted in a slightly better performance (e.g. Question 3, 4 and 5), whereas on questions 2 and 8 the responses to the unstructured questions used better strategies.

Where there were variations across the three levels of structure in the other dimensions - complexity, abstraction and resources, these tended to be a knock on effect of the differences in strategy. For example on Question 2 candidates taking the unstructured question scored higher

on the strategy dimension than other candidates. Students attempting the structured and semi-structured versions were uncertain how much was required of them in answer to the sub-questions. It also led to repetition and so created problems of time pressure which could account for the relative weakness of answers to (d) identified by the examiner.

The differences in complexity showed a similar trend with candidates doing the unstructured question better fulfilling the demands of the question. This was because complexity was reduced by the differences in strategy.

5.2.3.4 Summary of examiner's ratings of demands

As the examiner predicted, there was only small variance in the performance of candidates across the three versions. This reflects the statistical findings that structure did not have a significant effect on performance.

5.2.3.5 Demands found using the Repertory Grid technique

The repertory grid interview allowed the examiner to describe, using his own language, the constructs considered to constitute 'demands'. The repertory grid gave two areas of findings: firstly the constructs elicited by the examiner showed us what he considered to be the important features of demands in relation to the questions he looked at. Secondly, his ratings of structured, semi-structured and unstructured versions of two questions showed us how he rated each question on each construct.

	Characteristic of questions seen as demanding	Characteristic of questions seen as less demanding
1	No hints to what contributions/resources	Question indicated which area to flesh-out. Ideas were suggested to them.
2	Candidate must relate material to abstract concepts/ideas.	It is easier to introduce descriptive material (e.g. events and policies).
3	Candidate must evaluate material/ideas.	The question does not overtly require evaluation. The candidate can get a fair mark from description.
4	The candidate has to decide for themselves how to order/provide an argument for both sides.	The question provides structure for argument. It breaks down both sides of the argument.
5	Candidate is asked to do too many things in 45 minutes	They are given the opportunity to fit material into the time available.
6	The question forces them into repetition. Material is likely to be repeated because it is relevant to more than one part of the question.	No need to repeat because the candidate uses their own structure.
7	Candidate is left on their own to identify and sort out relevant material	They are led through the stages of the argument.
8	Candidate needs to be trained to conclude as a separate exercise (for a few marks). Candidates are unfamiliar with this questions type.	Candidates make judgements throughout and reach their own conclusion. This is a more familiar way for them to work.
9	Question parts are linked. Previous sub-question answers are needed to make subsequent judgements.	They are not required to make specific links - the candidate decides what the links are.
10	They are restricted in developing a different line of argument by the structure of the question.	They are given the freedom to structure the question as they see appropriate.
11	A brief answer is required to a question which could be lengthier.	They are given the freedom to decide how to allocate time to different arguments.
12	Required to balance complex judgements	Individual judgements are kept separate.
13	Interpretation of the focus of the question required of candidate	Focus of the questions is made explicit by wording.

5.2.3.5.1 *Relating the constructs to questions*

Six questions were rated on these constructs. The six questions were Question 1 (Henry VII: structured, semi-structured and unstructured) and Question 6 (Francis I: structured, semi-structured and unstructured) rated on a 5 points scale (where 1 was the extreme of the more demanding construct and 5 was the extreme of the contrast construct). Correlations between the ratings of the questions show us which questions were seen as most similar and which constructs were seen as most similar.

5.2.3.5.2 *Relationships between questions*

The questions most strongly seen as similar were:

Henry SS and Henry US.

Francis SS and Francis US.

This suggests that the question topic (Henry or Francis) was a better indicator of similarity than the level of structure (S, SS or US). This is further evidence for the statistical findings that the question topic was more influential on performance than the structure of the question.

Question	Factor 1	Factor 2	Factor 3
Francis Structured	-.73	.11	.53
Francis Semi-structured	-.40	.79	-.34
Francis Unstructured	.58	.65	-.43
Henry Structured	.09	-.79	-.46
Henry Semi-structured	.86	-.12	.19
Henry Unstructured	.72	.27	.53

5.2.3.5.3 *Relationships between constructs*

	Construct	Factor 1	Factor 2	Factor 3
1	**hints**	-.41	.83	
2	**abstract concepts**	-.47	-.67	
3	**evaluate**	-.54	-.42	.57
4	**structure argument**	.94		
5	**do many things**	.72	-.62	
6	**forced repetition**	.84	-.40	
7	**relevant selection**	.99		
8	**come to conclusion**	.92		
9	**linked parts**	.73	.46	.47
10	**restricted by the structure**	.95		
11	**brief answer**	.32	.49	.73
12	**balance complex judgements**	-.35	.54	-.67
13	**interpret focus**	.86	.32	

Factor analysis of the constructs showed that three factors emerged from the thirteen original constructs. These three factors accounted for 89% of the variation of the examiner's ratings.

Factor 1

4	The candidate has to decide for themselves how to order/provide an argument for both sides.	The question provides structure for argument. It breaks down both sides of the argument.
7	Candidate is left on their own to identify and sort out relevant material	They are led through the stages of the argument.
8	Candidate needs to be trained to conclude as a separate exercise (for a few marks). Candidates are unfamiliar with this questions type.	Candidates make judgements throughout and reach their own conclusion. This is a more familiar way for them to work.
10	They are restricted in developing a different line of argument by the structure of the question.	They are given the freedom to structure the question as they see appropriate.

This factor accounted for 39.9 % of the variation in the ratings. It pertains the degree to which the students organise their answer and present their argument. The examiner clearly saw this as a very important factor contributing to the demands of these questions.

Factor 2

1	No hints to what contributions/resources	Question indicated which area to flesh-out. Ideas were suggested to them.
2	Candidate must relate material to abstract concept/ideas	It is easier to introduce descriptive material (e.g., events and policies).
5	Candidate is asked to do too many things in 45 minutes	They are given the opportunity to fit material into the time available.

Accounting for 28.6% of the variance in ratings, this factor clearly relates to the identification of relevant historical material that is brought to the answer.

Factor 3

3	Candidate must evaluate material/ideas.	The question does not overtly require evaluation. The candidate can get a fair mark from description.
11	A brief answer is required to a question which could be lengthier.	They are given the freedom to decide how to allocate time to different arguments.
12	Required to balance complex judgements	Individual judgements are kept separate.

This factor accounted for 20.5% of variance in the examiner's ratings. This factor is about the need to link sub-questions and the difficulties candidates (who are used to responding to essay questions) have when having to respond to short answer questions.

5.2.3.5.4 Summary of Repertory Grid Interview results

Structure was not as influential on the examiner's ratings as the question topic.

Three main factors of demand emerged.

- **Demanding questions**

1. gave students opportunities to organise their answer and present their argument,
2. made students decide what historical material was relevant,
3. required that students (who were used to answering essay questions) respond to short answer questions.

5.2.4 ***Students' perceptions of demands***

Six candidates were interviewed and asked (i) how they tackle questions and (ii) their perceptions of the ease or difficulty of the three versions of the questions. The following themes emerged.

5.2.4.1 Specificity

This theme emerged from candidates' perceptions of how the question either restricted or freed them. Those structured questions which were more specific about the historical material to be included were reported to restrict students. These were thought to be on the one hand restrictive, yet on the other hand they were seen as potentially useful for candidates who could not recall the necessary material when they were under the pressure of the exam. One student preferred the unstructured question because he felt that 'you had more opportunities to bring in

extra knowledge and you could bring that in and make it relevant'. He attempted a structured version of question 7 (Isabella and Ferdinand), and had more to say in answer to part (b) than part (c) which were both worth the same number of marks. He would have preferred to have tackled the unstructured version, so that this discrepancy in his knowledge would not have been so obvious.

5.2.4.2 Organisation

This theme emerged from candidates' comments about the way in which the structured versions removed some of the need for them to organise their responses. One student described how the provision of a structure for organising their argument 'could undermine you'. She had devised a strategy for answering Question 1 by dealing with four themes. However, the structured version of the question (which she saw after planning her response) suggested the use of different themes. Clearly a high ability student, she was confident that her strategy was more effective than the suggested strategy as the suggested strategy which would have required repetition of material in the two sub-questions.

Students were concerned that if a strategy was prescribed for them then they would have problems allocating time to each sub-question. They thought that they would be more likely to run out of time or repeat information in following sub-questions.

Although the able students to whom we spoke preferred the unstructured questions they thought that the structured questions were probably easier. This was thought to be because 'it does the structuring for you'. Structured questions were described as '...more like GCSE than A Level'. Despite this, they did not perform any better on the more structured questions.

5.2.4.3 Expectations

All History candidates interviewed talked about the types of questions that they were expecting in their actual A Level exams and told us that these expectations highly influenced how easy or difficult they found the questions.

The A Level students (and teachers) had spent a significant amount of time and effort developing essay writing skills. They didn't know how to approach the short answer questions that appeared in the structured and semi-structured papers. They found it hard to write short answers rather than essays because 'writing essays is the way we've been taught to do it'. Because they were expecting, and had been taught to respond to essay questions, they found that it was difficult to decide what information to put into each of the sub-questions: 'when you are expecting an essay, but you get a question with only 5 marks it's hard to decide which information to put in which bit of the question.' One strategy used to overcome this difficulty was to approach each sub-question as if it were a mini-essay; 'You could structure them like an essay'. This strategy would cause them to repeat material across sub-questions which would 'eat in to the time'.

5.2.4.4 Validity

Interviewees were quick to point out where they thought different versions were not actually making comparable demands on candidates. For example the structured version of one question asked specifically for evaluation, whereas the unstructured version did not.

They were also concerned that the mark schemes should not be the same for the different versions because they saw the questions as really quite different.

5.2.4.5 Summary of interview findings

The interviewees identified three salient factors influencing what they found easy or hard:

1 the specificity of the question's historical content

2 the degree to which they were restricted to a given answering strategy

3 the preparation which they had undergone for responding to essay type questions.

They particularly saw the structured versions as restrictive and not allowing them the freedom to really demonstrate their ability. Their expectations of what type of question would arise and their preparation for those types of questions was a very important variable in their performance. This seemed to over-ride any other features of the question.

5.2.5 *The effects of structure on demands in A Level History*

Three sources of evidence (statistical, interviews with candidates and examiner's rating of demands) have been presented, each giving explanations for the finding that structuring of questions had no effect on students' performance. Two issues repeatedly arose:

1. The type of structuring used in the questions manipulated the strategy given to candidates to organise and present their answers (rather than giving prompts to relevant content). This type of structuring had no effect on performance.
2. Candidates are expecting essay questions, their teachers prepare them for such questions and thus they learn how to structure their arguments and present them in essay form. These expectations were a large contributor to their performance. Related to this issue is the expectations of the examiner. Like the students they have become experts in one particular type of question - the essay question. As the examiner commented after writing the questions and marking the responses:

   > "One was conscious that one was imposing on candidates an interpretation of the question. One was also aware that some essay questions are easier to turn into structured questions than others.... To put it another way, the real problem is setting good semi-structured and structured questions, a task in which we have no experience."

5.3 Combined Summary for GCSE and A Level History

5.3.1 *GCSE*

The investigation of GCSE History identified three factors which affected the demands of the questions.

1 The provision of prompts to relevant material.

2 The provision of scope for students to develop a strategy for responding.

3 The opportunity a question gave to use high level skills (such as making generalisations, synthesising, evaluating and analysing).

These were all mediated by the ability of the candidate attempting the question. Higher ability candidates were better able to cope (and the very able thrived) when presented with unstructured questions.

5.3.2 *A Level*

At A Level two factors emerged:

1 The most common way of structuring the questions was to provide a structure for candidates to organise and present their argument. This type of structuring did not influence the demands made on candidates.

2 The predictability of the question types at A Level allows candidates to fully prepare for the types of questions expected. Candidates know that they will be asked to respond to essay questions in their final exam. The fact that candidates are prepared to answer essay questions rendered any other type of question unfamiliar, and thus performance was poorer: this is the current situation. If the exam papers changed so would candidates' expectations and their preparation.

5.3.3 *The effect of structure on demands in History*

We can conclude that

1. Of the two types of structuring used [(i) prompts to relevant content and (ii) the provision of a strategy to organise and present responses] the latter of these had less effect than the former.
2. There were more opportunities in unstructured questions to show higher level skills. The more able students were better able to exploit these opportunities to their advantage than lower ability students.
3. Students' preparation and expectations were highly influential on their performance, in some cases having more impact than the supports given by structured questions.

6. SUMMARY OF SUBJECT FINDINGS

This chapter brings together the findings of the experimental phase of the project; it provides a summary of the findings for Geography, Chemistry and History, and finally general findings are presented.

To re-cap on the focus of the project, the research question asked:

'What are the effects of structuring exam questions on the demands made on candidates?'

For each subject there are three areas to consider: first, the effect that structuring had on performance; second, the effect that structuring had on demands and finally the students' perceptions of demands in structured and unstructured questions. A summary of each of these three areas is presented in this chapter.

6.1 Geography

6.1.1 *The effect of structure on difficulty*

The examiner wrote three versions (structured, semi-structured and unstructured) of each question. The different methods used by the examiner to change the structure affected performance by changing the cognitive demands of the questions in different ways. The following variations on structure were used:

1 The use of an everyday term in a technical sense.

2 The use of a difficult technical term.

3 Specifying the how the students' answers should be organised.

4 Providing cues to the content of an answer (changing the requirement from production to recognition).

5 The use of different command words such as 'describe' and 'explain', changing the nature of the answer required.

6 Breaking down the task into sub-parts.

7 Asking students to generate a theory rather than leading them into discovering it via a specific task.

The methods of structuring in A Level affected performance by changing the cognitive demands in similar but fewer ways:

8 The use of abstract and technical terms.

9 The use of prompts specifying how students should organise the content of their answer.

10 Dividing a question into two separate tasks rather than asking students to discuss two things at once.

11 The use of different command words, changing the nature of the task.

Structuring in Geography involved changing cognitive demands in a number of different ways, which resulted in the more structured questions being easier in most cases. Students performed significantly better on the structured version than the semi-structured and unstructured versions in two out of the four questions in GCSE, and in five out of the seven sub-questions in A Level.

6.1.2 *The effect of structuring on demands*

Examiners rated the questions on four types of demand (after Edwards and Dall'Alba 1981): *complexity* (of each component operation or idea and the links between them), *resources* (the use of data and information), *abstraction* (the extent to which the student deals with ideas rather then concrete objects or phenomena) and *strategy* (the extent to which the student devises or selects and maintains a strategy for tackling and answering the question). The questions in which structure affected performance had different demands in terms of *complexity* and *strategy* in GCSE; and *complexity*, *strategy* and *abstraction* in A Level.

Some of the cognitive demands introduced into questions were not intended by examiners, and these account for the cases in which semi-structured questions were more demanding than unstructured questions.

The examiners rated (i) the demands in the questions and (ii) the quality of responses. The responses to the structured, semi-structured and unstructured versions of the questions were rated differently, reflecting the different levels of performance on the three versions of the questions.

A repertory gird interview with the examiners, in which they described the demands in exam questions, demonstrated that there were three main types of demands in GCSE, these related to the demands of complexity, abstraction and structure/strategy. Four demands emerged at A Level which were complexity, resources, abstraction and structure/strategy.

6.1.3 *The students perceptions of demands in questions*

Students perceived demands in the questions to be different in the three versions at both GCSE and A Level. They thought the structured questions showed more clearly what the examiner expected, although these questions were also restrictive and less flexible than the unstructured questions which allowed them more opportunity for demonstrating their knowledge.

6.2 **Chemistry**

6.2.1 *The effect of structure on difficulty*

In Chemistry, candidates performed better on the more structured questions. At GCSE the effect was noticeable even at an individual question level, and accumulated over a whole paper could make a difference of at least 12 marks out of 100. Increasing the amount of structure

could increase the mean mark by the same amount. However, it should be emphasised that grade boundaries can be raised if overall marks are higher, so this would not necessarily result in higher grades. At A Level, the same effects were found as at GCSE, only to a much smaller extent.

At GCSE, six types of structuring could be identified, five of which affected performance.

1 Pure structuring: breaking the question into several parts without providing or requesting any extra information - *had no significant effect on performance.*

2 Varying the amount of information: usually better performance where more information was provided.

3 Greater specificity was usually associated with better performance.

4 Amount of language required: demanding a comprehensible English sentence or paragraph nearly always produced worse performance.

5 Restricting the response format: choosing a letter, ticking a box, putting a ring round a formula, or filling in a blank in a sentence, always generated better performance. (This may be due to the combination of two factors: gaining marks by chance from guessing, and the reduced demands of a recognition task rather than a production task.)

6 Giving clues to the strategy students should use was usually associated with better performance.

At A Level there were fewer types of structuring used. Most involved a combination of pure structure, information, and specificity. Providing extra information and increasing specificity usually, but not always, improved performance.

6.2.2 *The effect of structuring on demands*

Using the modified version of Edwards' scale of cognitive demands, most ratings given to all versions (whatever their level of structure) of the questions were at the low demand end of the scale. However, those questions which were seen as more demanding by the examiner tended to be the unstructured questions. Manipulations of structure altered the demands on the dimensions of *complexity* and *strategy*, but not *abstraction* or *resources*.

In GCSE Chemistry, there was no significant difference between the ratings of demands made in the question and the demands students were able to show in their response. The implication is that to test higher-order skills such as linking ideas and planning a strategy, less structured questions need to be set, since candidates will not be able to show high-level responses from low-level questions (see page 54: 4.1.2.1.)

At A Level, the repertory grid analysis of the examiner's perception of demand identified three types of demand, which could be described as 'familiarity', 'knowledge' and 'strategy'.

6.2.3 *The students perceptions of demands in questions*

Students found structured questions less intimidating and generally easier. They liked to have a breakdown of the mark allocation, and liked it to be made clear what the expected response to a question was. They recognised that some questions were more appropriate in an unstructured form, and that for a candidate on top of the material it might be easier to gain marks on this type of question. A Level students felt that structure did not really have that much effect since Chemistry is a subject where knowledge of the material is the decisive factor.

6.3 **History**

6.3.1 ***The effect of structure on difficulty***

There was no significant effect of structure on performance in any of the eight questions in A Level History, and an effect only for one question (of three) in GCSE History, where the semi-structured version was significantly harder than structured and unstructured version. This unexpected finding can be explained by the high ability of the sample of students taking that question. They were better able to cope with the open, unstructured question; they showed this by taking up the challenge of devising an effective strategy for answering and took up opportunities given to them in the question to show their higher level skills.

6.3.2 ***The effect of structuring on demands***

The investigation of GCSE History identified three factors which affected the demands of the questions. These were all mediated by the ability of the candidate attempting the question. Higher ability students were better able to cope (and the very able thrived) when presented with unstructured questions. The factors were:

1 The provision of prompts to relevant material.

2 The provision of scope for students to develop a strategy for responding.

3 The opportunity a question gave to use high level skills (such as making generalisations, synthesising, evaluating and analysing).

At A Level two factors emerged:

4 The most common way of structuring the questions was to provide a structure for students to organise and present their argument. This type of structuring did not influence the demands made on students.

5 The predictability of the question types at A Level allows students to prepare fully for the types of questions expected. Students knew that they would be asked to respond to essay questions in their final exam. The fact that students were prepared to answer essay questions rendered any other type of question unfamiliar, and thus performance was poorer.

The Repertory Grid technique showed the demands that the examiner considered to be present in questions. Two factors emerged from the GCSE questions: the first was the extent to which students were led through a question, by being given clues about organisation and relevant content; the second factor was the extent to which higher order skills (e.g. evaluating and concluding) were required. In A Level three factors emerged: organisation and presentation of answers, deciding on relevant material, and linking sub-parts of short answer questions.

6.3.3 *The students' perceptions of demands in questions*

Students' preparation and the expectations were highly influential on their performance, in most cases having more impact than the supports given by structured questions. The A Level students and the more able GCSE students preferred unstructured questions which allowed them more opportunity to show what they knew.

6.4 **Overall Summary**

In the three subjects, questions were structured in a variety of ways:

1 Questions were broken down into sub-parts;

2 The type of response format was varied.

3 The content was structured by giving prompts or extra information;

4 The answering process was structured by giving clues to strategies;

5 Language was manipulated;

All except 'pure structuring' (point 1) of these affected performance. However, in a subject where students were expecting essay questions, and were prepared to develop and maintain a strategy for answering, the type of structuring which gave students clues to the way they should answer did not improve their performance.

In Geography structured questions were usually easier than unstructured; in Chemistry structured questions were almost always easier; and in History students performed equally with different levels of structure.

Examiners found that questions of different structure had different demands in terms of the complexity of each component operation or idea and the links between them. They also had different demands in terms of the extent to which the student had to devise or select and maintain a strategy for tackling and answering the question. Differently structured questions were usually equivalent in terms of the use of data and information and the extent to which a student had to deal with ideas rather than concrete objects or phenomena.

Students preferred the sort of questions that they were used to and were prepared for. At GCSE this was either structured or semi-structured and at A Level it was invariably unstructured questions. In History the type of questions students were expecting was particularly influential on performance.

Conclusions and implications of these findings are discussed in Chapter 9.

7. THE EFFECT OF QUESTION STRUCTURE ON RELIABILITY

7.1 Introduction

One of the reasons often cited for introducing more structure into examination questions is a concern for reliability. It is supposed that increased structuring will mean more explicit questions with more predetermined, or more objective, correct answers. This is expected to increase the overall reliability of the assessment process.

It was not possible in this project to carry out a quantitative investigation to test the truth of these suppositions. To investigate inter-marker reliability in this way would involve large numbers of student responses being marked by at least two markers each, and would only indicate whether or not there was an effect caused by the manipulation of questions. To quantify the effect would require a wholly different design, with whole simulated examination papers, so that aspects of reliability other than inter-rater could be investigated.

This chapter describes the results of a qualitative empirical investigation into the phenomena associated with changes in question structure. The methodology used was described in Chapter 2: by eliciting comments and comparisons from examiners as they evaluated scripts the intention was to expose some of the nature of the judgement process. The findings should relate significantly to the issue of examination reliability.

One feature of the methodology needs specific comment. The comparison and judgement processes we investigated were experimentally controlled. In particular, the various versions of each question were intended to vary as little as possible *except* in terms of how structured they were. All three versions used the same content, and were marked using the same marking scheme. In real examining, if one of these formats were adopted, it is likely that some of the questions would have been altered more substantially to suit the format. Some questions seem naturally to fit one kind of format and might not be used at all if an inappropriate format were standard. This means that this study will underestimate the differences that would be seen in practice, between questions planned to be more structured or less structured. However, this also meant that whatever effects are detected here are themselves likely to be quite reliable indicators of the kinds we would see in reality.

7.2 Procedure

Five of the six components of the project were investigated in this study (practical problems prevented us from carrying out the procedures with GCSE History). Following the interviews, all comments made by the examiners were transcribed. From each transcript all comments that might relate to reliability, or that gave clues to the thinking processes of the examiner, were extracted and listed. These five lists were then combined, and the comments sorted according to the points that they seemed to be making. Twenty such points were obtained; they are reported below in three categories which relate to different aspects of reliability and validity.

In each section the comments are summarised, and then illustrated where appropriate with quotations from the examiners.

7.3 Reliability: consistency of marking

The initial concern over reliability relates to the unfairness that might be introduced to examining if different markers gave different marks to the same answer.

The study generated three clear conclusions relating to the general issue of the objectivity of examiners' marking.

- ***With structured questions examiners are more certain.***

Several examiners noted that they were not sure what mark to give a response, and this always occurred with responses to semi-structured or unstructured questions. For example the comments; "[Student P] may not have understood the question" and "[Student D] may have forgotten to answer the last part of the question" (both Chemistry GCSE) illustrate this. The A Level Chemistry examiner said he was "more likely to give *benefit of doubt* marks in the unstructured cases.

It is clear that reliability will be improved if examiners are more certain about the quality of the candidates' performance.

- ***With structured questions examiners demand more relevance.***

What was included and what was left out were both seen as very important in History. In marking responses to structured questions the A Level examiner would expect more relevance for the same marks. An 'unstructured' candidate who had missed out an important point "would now get an extra mark" when the A Level Geography examiner knew which version he had done, while a 'structured' candidate would lose a mark because his response was "less relevant to the structured question". He also said of one question that "in the unstructured version almost anything would be credited", and advocated semi-structured questions as best for differentiating at the top end.

To a considerable extent this leniency on relevance for unstructured questions is intended to compensate for the perceived increase in difficulty, but it is also clear that it places a greater burden on the examiner to judge how relevant some items are, and how lenient to be. We would therefore expect reliability to suffer.

- **With structured questions answers are more similar and more comparable.**

The "almost anything" goes comment quoted above relates to this. So does the GCSE Geography comment about the phrase *give an accurate location*. This was used in the unstructured version in place of the more precise *give the co-ordinates*, and the examiner noted that it "would require the mark scheme to be changed so that candidates could get more marks if they gave extra information", that is more information than the examiners really wanted. The A Level Geography examiner stated that "the unstructured version requires a more

permissive mark scheme". The A Level History examiner noted that candidates answering the structured version of one question "had been restricted in the answers they could give"; while this was meant as a comment in support of less structured questions in general it also suggests that less "restricted" responses would be more difficult to mark reliably.

Several other comments supported the idea that structuring generally reduces the outcome space available to students, giving them less freedom to show what they can do but also to avoid what they cannot. Whatever the impact on validity, structure in this way must improve reliability.

7.4 **Reliability: spreading out of candidates**

Reliability in practice, though, also indicates how successful an examination is in separating out the candidates, so that the final grade each one gets is likely to be stable across different parallel examinations. An interesting phenomenon arose here, which may depend on the degree of structure already present in current examinations as well as on the nature of the subject concerned.

In Chemistry a move to unstructured questions would **increase** the variance of marks.

In Geography a move to unstructured questions would **decrease** the variance of marks.

In History a move to structured questions would probably decrease the variance of marks, but would have more impact on validity.

7.4.1 ***Chemistry:***

The A Level Chemistry examiner "likes to spread them out", and is "happy" to give marks generously to those he thinks understand their Chemistry. It is clear that he found more to be happy about in the unstructured cases. The standard deviations of the marks he awarded indicate this (though they are based only on a very small number of cases):

	Structured	Semi-Structured	Unstructured
SD	0.43	1.30	2.55

The GCSE examiner noted that "there is not enough evidence in the structured version" to be sure of the candidate's ability, and other comments made it clear that both examiners would value some more open ended evidence of ability.

7.4.2 ***Geography:***

In contrast the Geography GCSE examiner said; "a higher level of response would now be required when headings were given to structure a question, and he later repeated this comment on another structured question. As already noted "almost everything would be credited" by the A Level examiner in responses to an unstructured question. The evidence from standard deviations at A Level is quite the opposite of that in Chemistry:

	Structured	Semi-Structured	Unstructured

SD	2.19	1.79	1.48

Both Geography examiners noted that credit would be given in unstructured questions for points that were really rather trivial, but these points would get no credit if given as part of the structuring of the question - there would be no 'free' marks.

7.4.3 ***History:***

It seemed that both of these views would apply in History, with structured questions both "restricting" the candidates and keeping them more "relevant". Perhaps this is why there was little variation in the spread of marks awarded on different types of question.

In summary, it is apparent that changes in structuring may have a substantial impact on the variance of marks awarded. Exactly what impact will depend on the extent of current question structuring and how that relates to the marker's need to see evidence of ability.

7.5 Validity: expectations

Although the study reported in this chapter was intended only to address issues of reliability, several of the comments recorded referred to candidates' expectations, an aspect of validity, and it seems worth reporting them briefly here.

Examiners, especially in History, were concerned that some of the question versions would be unfamiliar to candidates, or at least not be what they were expecting. Similar sentiments have already been noted from the interviews with pupils in Chapter 5. In this study the examiners were generally commenting on scripts without knowing how the different questions were constructed, and when they were shown the three versions it was common for them to remark on the differences in terms of what they thought candidates would expect to see in their examination paper.

We were left with a strong impression that teachers and pupils would quickly adapt to any change in format, to generate the sorts of answers examiners expect to see in the scripts. It seems very important that examination boards should make available to schools sufficient information about marking criteria for any open ended question - without this any mismatch between expectations and actual questions and mark schemes will constitute a serious threat to validity.

8. CHAPTER 8 STRUCTURE AND DEMANDS IN MATHEMATICS A-LEVEL

8.1 Structure

The experimental phase of this project involved the subjects Geography, History and Chemistry. In addition, however, a non-experimental study of A level mathematics was carried out.

In the main part of this project we have manipulated structure directly to investigate the effect on candidates' performance, and on experts' rating of demand. Here, however, there was no experimental manipulation. We compared a 1996 mathematics paper with a 1986 one, with a view to testing some of the assumptions above - namely investigating whether structure had changed, whether rated demand had changed, and (within the 1996 paper) whether any relationship could be found between the amount of structuring in the question and the level of performance as measured by the facility value.

The 1996 paper involved was paper 1 of the UCLES 9200 syllabus. The 1986 paper was paper 1 of the UCLES 9205 syllabus C.

8.1.1 Surface structure

In this section the structure of the two examinations will be compared in terms of what we have called 'surface' structure: rubric, paper layout, breakdown of questions into discrete parts, extra supports (diagrams and tables) and number of specific questions asked. (This means the number of answers, or pieces of information that the text of the question asks for, as distinct from intermediate stages in calculations etc which may be awarded marks according to the mark scheme.)

In the 1996 examination there were four papers. All candidates did paper 1, which was pure mathematics. On each paper the instruction to the candidates in the rubric was "There is no restriction on the number of questions you may attempt." This means there was no question choice as such, though candidates may have decided to spend their time answering less than the full number of questions.

In the 1986 examination there were two papers. Paper 1 was pure mathematics. The first section contained eleven compulsory questions, the second section allowed a choice of four questions from seven.

Both papers lasted three hours. In the 1986 paper there were 98 marks available; in the 1996 paper 120 marks were available.

The structure of paper 1 on the two examinations is summarised in the tables below. In these tables 'visible marks' refers to the mark totals for each question or part question that were printed in square brackets on the question paper. 'No. of explicit questions asked' is a count of the pieces of information candidates were asked to supply / calculate in each part - for example '... write down the maximum and minimum values of y and the values of t when they occur' was coded as asking for three pieces of information. There is obviously a certain amount of arbitrariness in these codings, but they are mainly straightforward. Structure of paper 1 in Mathematics A-Level 1996, syllabus 9200.

Paper 1 question	Total marks	Labelled parts i.e. a,b / i,ii	Unlabelled parts	Visible marks	No. of explicit questions asked.	Total number of lines of text	Other information
1	2			2	1	1	
2	3	2		3	3	2	diag
3	3		2	2,1	1,1	3	
4	4			4	1	2	
5	4			4	2	1	
6	4			4	1	2	diag
7	5		2	2,3	1,1	3	
8	5			5	1	1	
9	5		2	2,3	1,1	3	
10	6		2	5,1	1,1	3	
11	6	2		3,3	1,1	2	
12	8		2	3,5	1,1	2	
13	9	3		2,2,5	1,1,1	5	diag
14	9		2	5,4	2,1	7	2 diags
15	10		3	4,3,3	1,1,1	6	diag
16	11	4		3,3,3,2	1,1,1,1	5	
17	11		4	4,3,2,2	2,1,1,1	10	
18	15		3	4,5,6	3,1,1	6	

Structure of paper 1 in Mathematics A-Level 1986, syllabus 9205.

Paper 1 question	Total marks	Labelled parts i.e. a,b / i,ii	Unlabelled parts	Visible marks	No. of explicit questions asked.	Total number of lines of text	Other information
Section 1							
1	5			5	1	1	
2	4			4	1	1	
3	4			4	1	3	
4	4			4	1	2	
5	4			4	2	2	
6	5			5	1	2	
7	4			4	2	3	
8	4	2		4	1,1	2	
9	5			5	1	2	
10	4	2		4	1,1	3	
11	7			7	1	2	diag
Section 2*							
12	12	3		4,4,4	1,2,1	6	
13	12	3		6,6	3,3	6	
14	12	3		4,4,4	1,1,1	3	
15	12	2		6,6	1,1	3	
16	12		3	4,4,4	1,2,2	3	
17	12		3	2,5,5	1,1,2	5	
18	12		3	2,6,4	1,3,1	6	

* Candidates could choose any four questions from section 2

In 1996 the pattern was one of the number of marks available per question increasing from the start to the end of the paper (or section within the paper). In general, each part only asked one question, though there were cases where two pieces of information were asked for. The number of lines of text increased in line with the number of questions asked and marks available, as would be expected.

In the 1986 paper there were fewer diagrams, visible marks were less likely to be broken down, and more questions could be specifically asked within a part. However, at this surface level of structure, the differences between the two are small.

8.1.2 'Cognitive' structure

This is defined as features of the question which provide the candidate with a 'route' through the question, or in some other way affect the 'outcome space' of possible answers. The presence of such features is assumed to make the questions less demanding (than they could be, not than they should be!) almost by definition.

1. Structure which indicates correct strategy.

On the 1996 paper 1 the phrases 'by putting... / by using .... / by means of ...' occurred five times. These could be considered to be instances of direction to the appropriate method (or the method expected in the mark scheme), and hence a cognitive support.

One example was in question 3, which read:

Expand $\frac{1}{(1-x)^2}$, where $|x| < 1$, in ascending powers of x, up to and including the term in x^3.

You should simplify the coefficients. [2]

By putting $x = 10^{-4}$ in your expansion, find $\frac{1}{(0.9999)^2}$ correct to 12 decimal places.

[1]

In the second part of this question, it would have been possible to ask 'Hence find .... correct to 12 decimal places'. Supplying the value of x means that candidates are being shown exactly how the second part relates to the first part.

In addition to the five cases mentioned above there were four uses of the word 'hence', as in 'hence find...', or 'hence determine...'. These should perhaps not be considered as instances of cognitive support as such, more as ways of linking parts of the question together. However, they do nonetheless specify a route through the question for the candidate and could therefore be argued to reduce the amount of planning of strategy required. Two examples are given below:

Paper 1 question 5:

Express $\sin 4\theta$ in terms of $\sin 2\theta$ and $\cos 2\theta$, and hence express $\sin 4\theta / \sin \theta$ in terms of $\cos \theta$ only. [4]

Paper 1 question 12:

Express $\frac{3}{(2x+1)(x-1)}$ in partial fractions. [3]

Hence find the exact value of $\int_2^3 \frac{3}{(2x+1)(x-1)}\,dx$, giving your answer as a single logarithm

[5]

In both cases it may have been possible to ask the question following the 'hence' on its own, and expect the candidates to find the appropriate method themselves. However, this might have conflicted with assessment goals - e.g. assessing the ability to express formulae in terms of partial fractions, or the ability to apply the half-angle theorem, by denying candidates who could not select the right strategy (a difficult skill?) the chance to gain marks for something they could do. In other words, the questions are arguably more valid as they stand.

This last example shows that the level of 'structure' in a question is not independent of the validity of the question.

On the 1986 paper 1 there were five instances of the phrases 'by putting... / by using .... / by means of ...' (the same as 1996). There were three uses of 'hence', two of which were followed by 'or otherwise'. This is the only noticeable difference from the 1996 paper, where the phrase 'or otherwise' did not appear. Perhaps this represents a change in policy over the years in order to make mark schemes clearer.

2. Outcome space

A noticeable feature of the 1996 paper was how clearly it was specified what form the answer should take. On paper 1 the number of decimal places or significant figures required was specified seven times, and the form of the answer required in other cases was specified seven times. (e.g. ' ... giving your answer in the form $ax + by + c = 0$...'). This form of structuring reduces ambiguity and presumably makes the marking easier, and the test more valid.

On the 1986 paper the number of decimal places or significant figures was specified twice, as was the form of the answer required. This could be taken as evidence of less structuring, but it may be due to the greater symbolic (rather than numerical) nature of the questions in 1986 - there was perhaps likely to be less confusion about what form the answer should take.

In conclusion, there was very little difference between the 1996 paper and the 1986 paper in terms of structure.

8.2 Question Difficulty

This project is looking at the relationship between the structure of an exam question and its 'demand'. One of the more interesting features of the project has been the attempt to determine what exactly is meant by 'demand'. It is apparently something separate from the 'difficulty' of a question, which is defined operationally here as being measured by the 'facility value' (the mean mark on a question expressed as a proportion of the maximum mark available - the lower

the facility value the more difficult the question). 'Demand' seems to be conceptualised more in terms of the skills required by the question, and is generally a more abstract concept. A detailed discussion of demand can be found in the main report. However, almost by definition we would expect the difficulty of a question to be related to its demand. In this section the difficulties of the questions on the 1996 paper 1 are presented. (There is no data on item level performance from the 1986 examination to compare it with). These data come from a sample of 200 candidates. The table on page 107 is taken from an UCLES internal evaluation report[1]:

One feature that can be seen is that, in general, where there is more than one part to a question, later parts are more difficult than earlier parts (have a lower facility value). This suggests that examiners are succeeding in structuring questions by providing an incline of difficulty. The same effect appears between questions - i.e. later questions are more difficult than earlier ones, though the effect is not so strong. The reason is that within a question, the answer to a later part may depend on the earlier part, as in the 'hence...' questions discussed earlier. Thus although mark schemes often allow 'follow through' marks from an earlier error, it would still be expected that overall the success rate on the later part would be lower. There is no clear relationship between the difficulty and the presence or absence of a 'cognitive support', but since these supports are present and absent on different questions, there is no relevant comparison.

[1] *'A-Level Mathematics (9200) June 1996: an evaluation of the masurement characteristics and quality of the examination.'* Nick Raikes 1997

Mathematics (Syllabus 9200) 1996, Paper 1 - item statistics

Item	Max mark	Facility value	Cognitive support ?
Q1	2	0.92	
Q2	3	0.81	
Q3_i	2	0.68	
Q3_ii	1	0.41	By putting...
Q4	4	0.45	
Q5	4	0.58	Hence..
Q6	4	0.68	
Q7_i	2	0.46	
Q7_ii	3	0.63	By using...
Q8	5	0.70	
Q9_i	2	0.68	By means of
Q9_ii	3	0.62	Use...
Q10_i	5	0.67	
Q10_ii	1	0.30	Hence...
Q11_a	3	0.89	
Q11_b	3	0.86	
Q12_i	3	0.91	
Q12_ii	5	0.72	Hence...
Q13_i	2	0.91	
Q13_ii	2	0.51	
Q13_iii	5	0.59	
Q14_i	5	0.89	By finding...
Q14_ii	4	0.54	
Q15_i	4	0.84	
Q15_ii	3	0.80	
Q15_iii	3	0.79	Using...
Q16_i	3	0.60	
Q16_ii	3	0.59	
Q16_iii	3	0.38	
Q16_iv	2	0.59	
Q17_i	4	0.45	
Q17_ii	3	0.38	
Q17_iii	2	0.22	
Q17_iv	2	0.28	
Q18_i	4	0.69	
Q18_ii	5	0.39	
Q18_iii	6	0.22	Hence...

8.3 Demand

Previous work (e.g. SRAC 1990, chapter 3)[2] has found that it is extremely difficult to explicate the concept of demand. The term is not always used consistently - it is sometimes interchanged with 'difficulty'. The SRAC paper suggested a breakdown of 'demand' into 3 categories:

- 'academic demand' - the intrinsic level of difficulty
- 'contextual demand(s)' - which relate to the surface structure and cognitive structure described earlier - i.e. the rubric, layout, mark allocation, amount of help offered in working through the question etc, as well as other demands like time pressure and the need to cover the syllabus.
- 'personal demand' - commitment, motivation, liking for / familiarity with topic.

The overall demand is an interaction of these idiosyncratic and unmeasurable demands. In the words of the SRAC study: "the judges ... were able to recognise demand when they saw it, and, in general, agreed with each other in categorising questions, papers and examinations according to their overall level of demand, yet they found it very hard to identify what construed that demand, to explain why they found something demanding."

That study was based, in part, on an earlier study of university mathematics questions which was carried out by Griffiths & McClone (1979, 1984). They devised several qualities (not necessarily demands as such) which judges rated questions for, using a 0 - 3 scale. These were:

1. Procedure - degree to which method of solution is open to the student
2. Objectives - degree to which question explicitly identifies the conclusions of the solution it is seeking
3. Jargon
4. Mathematical content apart from jargon and bookwork - (obviously somewhat arbitrary)
5. Definition, bookwork, stock example
6. Abstraction - closeness to theory, compared with applications remote from that theory
7. Mathematical manipulation
8. Logical manipulation - powers of reasoning required
9. Sustained thinking
10. Open solution - how clear it is what would constitute a solution.

These classifications were used to compare the questions set in different areas of study (i.e. mechanics, statistics, computing etc) at 10 different universities. For example, Computing was shown to be assessed by more open questions in terms of both procedure and objective (1 & 2

---

[2] *'A study of the demands made by the two approaches to "Double Mathematics"'* An investigation conducted by the Standing Research Advisory Committee of the GCE Examining Boards, published by UCLES in 1990

above), and to contain more jargon. It also allowed comparisons between universities in terms of how much each differed from the overall pattern - for example "University II overall places more emphasis on concrete rather than abstract questions and develops a more open approach in terms of procedure and objective."

This kind of detailed analysis is probably not so appropriate with the more uniform style of questions in an A-Level paper, but it illustrates how questions may be conceived in different dimensions, which again would all interact to affect the difficulty (or demand) of the question. In fact, the same dimensions were used in the 1990 SRAC study, but they reported that the exercise was "designed to assist in defining more closely the factors which contribute to demand as represented by individual questions in the examination. It was not intended as a basis for carrying out any sophisticated statistical analysis, and none was undertaken. Indeed the subjective nature of the exercise would make any conclusion from such an exercise somewhat questionable."

In summary, the earlier work has shown that while it is possible to distinguish many factors which make up the concept 'demand', it is not possible to specify in great detail (or any detail!) how they interact to influence the difficulty of a question.

Comparison of rated demand in 1996 with 1986

For the purposes of this study, we have rated questions using a modified version of the Edwards scale of cognitive demand (see chapter 1).

An A-Level examiner rated each of the 18 (whole) questions in the 1986 and 1996 papers on the dimensions of Complexity, Resources, Abstraction and Strategy. The results are summarised in the frequency table below.

	Rating	'2'	'3'	'4'	'5'
1986	Complexity	0	3	15	0
	Resources	13	5	0	0
	Abstraction	10	5	3	0
	Strategy	6	6	5	1
1996	Complexity	0	2	16	0
	Resources	13	5	0	0
	Abstraction	11	6	1	0
	Strategy	6	7	4	1

N.B. Although there were 18 questions on each paper, in the 1986 exam there was a choice of 4 questions from 7 in section B, whereas in 1996 it was possible to do all 18 questions.

Inspection of the table shows virtually no difference at all in rated demand between the two years, on any of the four dimensions. Remember that the 'year' of the exam represents the manipulation of structure here, the assumption being that the 1986 exam was less structured. As the earlier section showed, there was in fact little difference in structure between the two exams. Of the four dimensions, 'Strategy' seems the one that is most capable of discriminating among different A-Level questions.

8.4 Conclusion

In the context of comparing a 1996 paper with a 1986 one in terms of structure and demand, we found very few differences in either structure or demand. This may have been because we were comparing two examinations from the same board, but this seems to be a reasonable comparison as far as it relates to providing information on standards over time. The discussion of use of structure showed that its main achievement is to increase the validity of the questions rather than affect their demand.

9. CONCLUSIONS

The focus of this investigation was to discover the effects of structuring questions on the demands those questions made on candidates. Overall we have found that:

- In geography and chemistry, more structured questions usually made fewer demands on students and performance was better
- In history, the ability and expectations of the students overrode any effects of structure

Five types of structuring were used by the examiners to produce structured, semi-structured and unstructured versions of the same questions. Not all of these affected the performance of candidates or the demands made on them. There were various effects of structuring, depending on the subject, the type of structuring, the question types students were expecting, and the ability of students.

9.1 Five types of structure

1 ***Pure structure.*** Questions were structured by dividing a single sentence into a number of discrete parts, keeping the content of the questions identical.

2 ***Response format.*** Questions were structured by varying the way in which the student was required to respond.

3 ***Content.*** Examiners structured the content of a question by giving pointers to relevant material which could be included in an answer.

4 ***Process.*** Some questions were structured by specifying the process or strategy which a student could use to answer the question.

5 ***Language.*** Some examiners manipulated structure by varying the language used in the question.

These five types of structure are discussed below in terms of their effects on performance and demands made on students.

9.1.1 Pure structure

Effects of pure structure

Questions structured in this way asked for the same information in several discrete parts as the matching unstructured question asked in a single sentence. This type of structuring had no effect on performance, showing that students were not affected by structuring the surface features of a question. It seemed that this kind of structuring only changed the appearance of the question and did not change the cognitive processes necessary to answer the question. The content of the differently structured questions was the same and the processes required to answer them were also essentially the same.

Students' comments

For questions that were broken down into sub-parts, geography students commented that they used the mark allocations to organise their responses, in terms of how much to write for each question, and also to organise their time. Chemistry students used the breakdown of marks to decide the level of detail in their answers, although the negative side of breaking down the question was that it took longer to read.

Examiners' ratings

When a question was broken down into sub-parts, examiners rated this as having fewer demands in terms of strategy. Although there was no statistically significant difference in performance with this type of structuring, the examiner would have expected a difference to occur.

Implications

Breaking down written questions into sub-parts had no effect on performance. However, a situation in which pure structure may affect cognitive demands in an important way is, for example, in an oral question such as listening comprehension. Written questions are available throughout the answering process, but oral questions are not, and students have to store information in working memory. It may be desirable to reduce demands made by a particular question on working memory processes. Breaking the question down into sub-parts whilst keeping the content constant is a suitable method for doing this - at least in principle.

9.1.2 Response format

Effects of structuring the response format

Another way in which the examiners structured questions, particularly in chemistry, was by varying the format of the response required. For example, responses to the structured questions necessitated less extended writing, and instead involved ticking boxes or choosing an answer as in multiple choice. Candidates performed better on questions with a very structured response format. Choosing the correct answer from given alternatives tested recognition skills, whereas composing a written answer required production skills. The structured response format also reduced demands on linguistic skills, which has implications for validity. In science questions, for example, it may be desirable to reduce linguistic demands in order to test subject-related skills.

Students' comments

The response format is usually very predictable and a large amount of teaching and learning goes into preparing students to deal with the expected response format. An unexpected response format can cause confusion. Chemistry students felt that structured response formats were an advantage as there was no need to construct a sentence.

Examiners' ratings

The chemistry examiners rated questions with simple non-linguistic response formats as having very low demands in terms of complexity and strategy.

The examiners were all experienced in writing and marking certain types of questions, and had difficulty writing questions at levels of structure unlike the exams for which they develop questions.

Implications

A structured response format results in a smaller outcome space. The likelihood of overlap between the problem space defined by the examiner and the outcome space produced by the student is greater when the type of response is restricted. This contributes to validity, as the students' responses are likely to be more valid representations of their knowledge and understanding of the topic when the problem space is clearly defined. The responses can be

more easily related to the mark scheme, which increases the consistency of marking. The use of different response formats therefore has implications for both the validity and the reliability of assessment.

9.1.3 Content

Effects of structuring content

Examiners structured the content of a question by giving pointers to material that could be included in an answer. Providing this information had the effect of changing the task from uncued to cued recall. This reduced the cognitive demands of the question, as cues can aid the retrieval of information from memory (Tulving, 1983). In general, this kind of structuring improved performance significantly. Structuring the content of the answer may also have the effect of structuring the answering process as clues to relevant content may be in the form of prompts or sub-headings that provide support in devising a strategy (see section 7.1.4 below).

Students' comments

When headings were given, geography and chemistry students felt they knew exactly what the examiner wanted to read, and without these they may have missed out some of the relevant content. This means that responses to a question without headings may not be valid representations of students' knowledge, as students may miss out something that they knew but did not realise was relevant. In history a number of students commented that they did not need the prompts, but were reassured by them and would find them useful if they had recall problems induced by exam nerves.

Examiners' ratings

When cues to the content of an answer were provided in geography, examiners rated questions as having fewer demands in terms of strategy, as this changed the requirement from production to recognition. When students were told what to focus on in a diagram, they were able to come up with a theory, whereas examiners noted from the responses that students found this more difficult in the less structured question. Chemistry examiners rated questions in which students were given information as being less demanding in terms of strategy than ones in which students used their own knowledge. The history A level examiner was cautious about giving prompts to content in questions because they may restrict the more able candidates, and this

went against the aim of the syllabus and the examination to allow students to demonstrate their abilities. Indeed, it was found that the more able students preferred questions without prompts which allowed them freedom to include the material that they saw as relevant.

Implications

Structuring a question by providing information about relevant content specifies the problem space and also the outcome space by indicating what should be included, and this will increase validity in the sense of keeping the responses relevant. The likelihood that what the student writes matches what the examiner expects is greater with this type of structuring. It keeps students on task and can therefore give a better indication of their level of understanding. A problem space that is not clearly defined results in a large outcome space which means that students will be more likely to include irrelevant material. However, headings that indicate relevant content can cause difficulty if they are unexpected, and can also be restricting.

Removing all supports such as headings is a risky strategy. There is the potential for gaining a great deal of information about students' knowledge but also for gaining nothing at all. An open-ended question allows those candidates who have the skill to write an essay and form a coherent argument to show that they can do this. However, those who do not have the necessary writing skills would not have the opportunity in this kind of question to show what they do actually know about the topic. Their knowledge would be masked by the demand to express it in a certain way. The task would therefore be less valid than a structured task for those students. This problem could be avoided by using differently structured questions in different tiers in those GCSE examinations which allow tiering.

Changing the demands from uncued to cued recall changes the nature of the task, and again this may or may not be thought appropriate in different situations. In some cases it may be valid to test recall, but in others it may be important to discover the knowledge level of students without confounding this with their recall abilities.

9.1.4 Process

Effects of structuring process

In geography and chemistry, questions which provided candidates with a strategy (or clues to a strategy) for answering resulted in better performance than questions which required students

to develop and monitor their own strategy. In history however, one of the questions showed that providing a strategy can affect performance in the opposite way. In one of the three questions in GCSE history, students performed best on the unstructured version. By looking at the predicted grades it was found that the students answering all versions of this question were of high ability. They seemed to find the imposed structure in this question restricting as it did not correspond to the strategy they had been taught to use. In the other two questions there was no effect of structure, and the students were of mixed ability. Even though providing a strategy may have helped lower ability students in history, this effect could have been masked by the higher ability students performing poorly on structured questions as they were highly trained to answer questions according to a particular strategy which did not match the one given.

Giving students a structure to follow removed the need for them to monitor their strategy. For example, in chemistry a question involving a calculation was broken down into three parts specifying the three stages of the calculation. In geography, the answering process was sometimes structured by changing command words. When students were asked to describe rather than explain something, a different, less demanding answer was required, and performance improved. Certain command words can result in students writing a 'textbook answer', that is reproducing an answer that their teacher has taught them.

Students' comments

Students doing geography felt that being given a strategy prevented them from going off at a tangent. They felt that this type of structuring was best at the start of a question, to lead them into it. Chemistry students also saw the value of being given the organisation of their answer. In history A level, the high ability students who were given a strategy to follow felt this to be restrictive, especially as they were trained to answer unstructured questions.

Examiners' ratings

In geography, the examiner rated questions and responses as differing in demands in terms of strategy when the questions specified how students should organise their answers, and when command words such as describe and explain were varied. Chemistry examiners also rated questions in which process was structured as having fewer demands in terms of strategy.

History questions gave students 'opportunities' to show their ability, and the opportunities - that is the demands - in structured questions were described as fewer than those in the unstructured questions. Changing the way that students would tackle the GCSE history questions was seen as a problem because the types of questions which they were expecting were of a specific level of structure.

Implications

Provided the strategy given corresponds to the strategy taught, most students will perform better on a question in which the answering process has been structured. The problem space is reduced when the strategy required by the examiner is indicated in the question. This type of structuring may be valid in some but not in all cases. For example, in history the development of an argument is an important skill and it may be invalid always to remove this requirement by providing students with a strategy. One of the objectives of the A level history assessment was to test "the ability to present a clear, concise, logical and relevant argument." (OCEAC (1996) pg 4). Similarly in geography and chemistry, at least at A level, developing and monitoring a response strategy may be a valid subject-related skill.

However, a question that is unstructured in terms of the answering process can result in students giving textbook answers, that is repeating a standard answer that they have been taught, with little reference to the question set. A slightly more structured question can eliminate this possibility by requiring that students apply their knowledge in a more focused manner, but a very structured question would again be less demanding. It can often therefore be the semi-structured question that is the most demanding as it can require application of knowledge rather than regurgitation of an essay or a string of facts.

9.1.5 Language

Effects of structuring using language

In geography in particular, the examiners structured questions by changing wording. Abstract and technical words were used in the less structured versions, with everyday language being used in the more structured questions. This type of structuring introduced difficulties that were not intended by the examiners (see Fisher-Hoch et al., 1997 for details of potential sources of difficulty). Changing the wording can make a question ambiguous and misleading (although it can also improve clarity and understanding). If a student does not understand a particular technical term the question can become inaccessible. This type of structuring should therefore be used with caution.

Students' comments

Geography students recognised that the use of more geographical words tested their knowledge better than the use of everyday language, although they did find this hard. Syllabuses often outline the essential subject specific terms that should be known by students, and this provides a guide for teachers and question writers.

Examiners' ratings

Examiners in geography rated questions as more complex when technical vocabulary was used. However, they did not identify this as having affected levels of demands shown in the responses.

Implications

Students' use and understanding of technical terms is often required and is therefore valid. Using technical language allows examiners to test students' understanding of the questions. However, the use of everyday rather than technical language in questions could make them more accessible to those who do not understand the technical vocabulary (lower tier candidates for example). If these students cannot gain access to the question to even attempt it then it is an invalid way of assessing their knowledge. One way of dealing with this would be using questions in which the first part asks for the meaning of a technical term and the rest tests

understanding of the concepts involved in the topic. Ideally examiners could test understanding of the term and then give the meaning for those who have not understood it so that all students can go on to the rest of the question. This may be possible in the future with the use of sequential testing on computers. If students are not allowed to return and correct previous answers then examiners can discover whether they understand the terminology and then allow all students to demonstrate their understanding of the topic.

9.2 Discussion

It is not necessarily the case that structured questions are easier than unstructured questions. Changes in performance which arise through the structuring of a question can be affected by the nature of the structuring; by the expectations of (i) the teachers and (ii) the students (i.e. the predictability of questions), (iii) by the experience of the examiners and (iv) by the ability of the students.

9.2.1 Expectations

9.2.1.1 Teachers' expectations

Students are encouraged to see past papers as a prime source of information about the content of an exam. Teachers have access to detailed criteria for assessment through the past papers, published mark schemes, syllabuses, examiners' reports as well as INSET provided by the exam boards. A mark scheme may provide an implicit structure for an unstructured question. All of this information affects teaching, so that students are prepared for the types of questions that are expected to occur in particular exams.

In UCLES History A level (9020) candidates sit 2 out of 18 papers, and attempt 3 out of 25 essay questions on each paper. Students also choose a document based question on each paper. Because of the many options, questions within and across papers have to be comparable and this can be achieved by specifying the types of questions that will be asked. The syllabus states that:

> "In order to provide some predictability for candidates and teachers, and also guidelines for setters and moderators, the structure of questions which follow the

> documentary extracts will contain 4 or 5 clearly separate sub-questions...."
> (OCEAC 1996, pg 6)

Similarly for MEG History GCSE (1607) specimen and past papers show the types of questions which can be expected, as does the syllabus.

Some written comments from teachers whose students participated in this study showed that the students expectations were highly influential on performance:

> "... their familiarity with the unstructured format and their analytical training did probably prejudice them."
>
> "Staff who administered the papers (A level History teachers) felt that a poor answer in the structured questions might in some way be attributed to the fact that students have been trained to answer straightforward essay questions."

The phrase 'straightforward essay questions' illustrates how important expectations are. These students are taught exactly how to answer the types of questions they are expecting, and they are able to produce the appropriate style of answer in an exam. A more structured question can be challenging as students cannot now reproduce what they have learned, but have to think about the question and apply their knowledge to specific examples. This indicates that the types of questions the teachers are expecting has an influence on the types of questions the students are prepared for. Communication between examiners and teachers about the level of structure in questions therefore affects students' preparation.

The GCSE geography papers used in this study were based on MEG Geography GCSE (1588), and the syllabus for this course gives some indication of the level of structure in questions. There are two tiers, and most of the questions across tiers are based on common resources and content. Differentiation is therefore achieved by structuring to some extent:

> "...stepped questions will be set with an incline of difficulty across the sub-sections of each question. The earlier sub-sections will require short answers while the last sub-section will require a response in extended prose... .Foundation Tier questions will be structured to allow full marks for a short answer and a framework will be provided in the answer booklet for any extended writing. Higher Tier papers will be answered on plain paper so that the candidates can extend their answers as their ability allows."

Structuring is therefore used to make questions more accessible to lower tier candidates, and the way in which questions will be structured is presented clearly to teachers so that the students will know what to expect. The UCLES Geography A level (9050) syllabus also provides detailed information about the structure of the questions on each paper.

The MEG Chemistry GCSE (1781) syllabus simply states "Structured questions will be used across all question papers." The UCLES Chemistry A level (9254) syllabus indicates that there will be structured questions in Paper 2, without going into any more detail.

9.2.1.2 **Students' expectations**

As well as the expectations of teachers influencing what students learn, students themselves have expectations and ideas about question structure.

Students doing the geography papers perceived the demands in the questions to be different across the three differently structured versions. Structured questions showed more clearly what the examiner wanted. However, students found some structured questions restrictive and inflexible. Unstructured questions allowed them the opportunity to demonstrate their knowledge, and also allowed them to write about what they did know without revealing gaps in their knowledge.

Students doing the chemistry papers correctly perceived the structured questions as easier than the unstructured, and also as being a clearer indication of what the examiner wanted, although these students also felt that structured questions could restrict them. It was thought that the semi-structured questions were a good compromise.

Students doing the history papers said that they thought the structured or semi-structured questions would be easier, but they preferred the unstructured version which was closest to the type of questions with which they were familiar.

9.2.1.3 **Examiners**

The examiners writing the geography questions had some experience of writing questions at different levels of structure as MEG GCSE Geography (1588) is tiered in this way. However,

these examiners attempted many different types of structuring, sometimes changing wording, sometimes giving prompts to relevant content or to the answering process. Some of these methods produced the desired effects but some produced misleading questions and changed demands in ways that were not intended.

The chemistry examiners were not experienced in writing questions at different levels of structure and found it difficult to produce the unstructured questions.

In history, differentiation is by outcome and there is no tiering. The examiners were therefore not familiar with the process of structuring a question. The history A level examiner who has always set essay (unstructured) questions for the live exams summed this up:

> "... the real problem is setting good semi-structured and structured questions, a task in which we have no experience."

In general, all of the examiners preferred to write questions in the way they were used to, and found it hard to write questions at a different level of structure. The result of traditions of question writing is that examiners are highly experienced in writing certain types of questions and therefore produce very good questions. However, this also means that they find it difficult to change question structure, and may also be unwilling to do so as it can mean not writing as good a question as they could.

9.2.1.4 Ability of students

The effect of structuring in history was found to interact with the ability of students, as indicated by their predicted grades. This is an area which merits further investigation. Structuring may affect different students in different ways. The student who has limited knowledge of the topic may be prompted to recall some of this knowledge in a structured question, and may gain some marks by being able to do part of it, as long as the way in which the question is structured matches their expectations, and corresponds to the part of the topic which they know something about. In an unstructured question this student may not have any starting point from which to access the question. Students who know a great deal about the topic area may find structuring useful as it can focus them into including only the relevant aspects of their knowledge. However, without structure they may go off on tangents, including everything that they know about the topic but not necessarily the points required in the

question. It may be that only the very able have the flexibility to select relevant information from their knowledge base and apply it to an unstructured question. However, unstructured questions also allow those who have gaps in their knowledge to write a response that includes what they do know, without revealing the gaps in their knowledge that may be revealed in a structured question.

9.3 The Effects of Structure on Demands

Structuring exam questions has both benefits and disadvantages. Some aspects of structuring can fall into both of these categories, being beneficial in some circumstances but disadvantageous in others, as indicated in the table below.

Structure has been introduced into some exams in order to reduce non-subject demands such as linguistic demands, and to make exams more accessible as the ability range of candidates has extended. Structuring can be used to make questions more accessible, but this reduces the discrimination powers of the exam. For example in some of the GCSE chemistry questions the mark range was greater in the unstructured than structured versions. The unstructured versions spread out the students' marks, whereas the more accessible structured versions resulted in most students scoring high marks. As well as the dangers of making questions too accessible it is also important not to make them too demanding. A question that is both accessible and demanding can cause marks not to be evenly distributed across the mark range, but instead to be bunched around the middle. The effect of this is that grade boundaries are not far enough apart and this should be kept in mind when demands are changed. A change in question structure resulting in a change in demands may also therefore require a change in the mark scheme. One way of avoiding this is to confine structuring to lower tier papers.

Benefits of Structure		Disadvantages of Structure
Improves communication between examiner and student about exactly what is required (as long as technical terms are understood)		
Allows more specific mark schemes, increasing the reliability of marking		
Mark allocation for sub-parts helps students allocate their time		
Prevents students from rehearsing whole answers		
Examiner can choose to increase complexity incrementally throughout a structured question so that it is both accessible and demanding		
		Contains more reading and can look daunting
Reduces the need for students to develop and monitor an answering strategy	☒	Restricts opportunities for students to show they are able to develop and monitor answering strategies
Reduces non-subject demands, such as the linguistic demands of extended writing	☒	Can reduce demands which are desirable
Reduces demands on working memory	☒	Restricts opportunities for students to show they are able to recall relevant material
Makes questions accessible to lower ability students	☒	Reduces discrimination of questions

As students at GCSE level usually do a wide range of subjects, it is not crucial to have a range of structure within one exam. It may be more appropriate for some subjects to have structured questions and others to have unstructured questions, as most students will then gain experience of both types of question. Various aspects of students' skills will then be tested in different subjects. At A level it becomes more vital to have exams with a range of differently structured questions as students are more specialised in their choice of subjects.

Unstructured questions allow examiners to test students' ability to write and produce coherent arguments, whereas structured questions allow examiners to test specific subject knowledge in a more controlled way. The usefulness of differently structured questions therefore depends on the aims of the exam.

9.4 Summary

As detailed in chapter 1, structure was introduced into school exams for three reasons: to increase validity by reducing non-subject demands; to increase reliability of marking and; to control the difficulty of questions. A number of implications of structuring questions have arisen from this project. Five of them have important implications for reliability and validity.

9.4.1 Higher marks gained in structured questions

Generally, candidates performed better on structured questions than open questions. However, higher marks need not mean that it is easier to gain a higher grade (if that was the case then standards would be threatened). It is simple to adjust for higher marks when placing grade boundaries.

9.4.2 Expectations of students and teachers

The size of the problem space and outcome space is reduced when students and teachers are expecting certain types of questions. Reduction in problem space and outcome space increases reliability of marking, because the number of possible answers is limited. When students are well prepared for the type of question which will arise they are not being assessed on test skills, but on the subject content test validity will be higher. The implication is that any changes in question type should be gradual with support for the teacher (and thus pupils) as well as examiners.

9.4.3 The relevance of non-subject demands.

Structured questions can reduce non-subject demands (for example extended writing, using a range of vocabulary, and critical thinking). This can be a useful way of increasing the validity of questions, if these demands are seen as invalid: if they are valid they could be specifically targeted. Structure can allow students to get straight to the subject demands, without having to negotiate the potential barriers of non-subject demands.

9.4.4 The use of technical terms.

Technical language in examinations can make questions inaccessible and cause a barrier for those candidates who do not understand the terminology. On one hand, if students cannot gain

access to the question to even attempt it then this is an invalid way of assessing knowledge. Yet, on the other hand in many cases it may be desirable to assess students' knowledge of technical language, and testing such knowledge would then be valid.

9.4.5 The influence of ability on the effect of structure.

A spread of scores is necessary for assessment to be reliable. Getting a balance between accessible questions and demanding questions is therefore central to reliable assessment. Tiering has been used to provide the right balance of demands, since it enables us to match the style of the questions better to the ability level of the candidates. There is a need for questions to be accessible enough for the lower ability students to show their abilities, yet demanding enough to challenge higher ability students. But there is a danger that if we attempt to make a common paper *both* accessible *and* demanding the result may be a dramatic reduction in the effective mark range and a concomitant reduction in reliability.

GLOSSARY OF TECHNICAL TERMS

Abstraction	Dealing with ideas rather than concrete objects or phenomena.
Complexity	The use of complex ideas that are linked with one another.
Cued recall	Remember something that is not given, but is prompted.
Demands	Tasks within questions.
Discrimination	A test discriminates well if candidates' scores are spread over the full mark range available.
Outcome space	The set of all candidates' solutions to the problem both correct and incorrect.
Problem space	The set of all intended routes from the initial representation of the problem to the solution.
Recall	Remember something that is not given.
Recognition	The ability to remember something that is given.
Reliability	A reliable test is one that is consistent with other tests of the same thing, and would produce similar scores when repeatedly testing the same people.
Repertory Grid Technique	Kelly's (1955) interview technique, in which interviewees are given groups of three items (e.g. exam questions) and asked to say why one is different from the other two. It is an attempt to elicit constructs (in this case descriptions of demands) that are otherwise implicit.
Resources	The use of data and information given in the question.
Strategy	Devising or selecting and maintaining a method for tackling a question.
Validity	This can be taken to mean construct validity in this report. It is the suitability of the exam or question for testing the construct it is trying to test. For example a question is a valid test of geography if it requires students to demonstrate geographical knowledge, but not if the geographical element of the question is minimal and other skills are tested.
Working memory	A system for storing information that is currently in use, and manipulating it for use in other cognitive processes.

REFERENCES

Ausubel D P, Novak J D & Hanesien H. (1978) *Educational Psychology: A cognitive view* 2nd edition, Holt, Reinhart and Winston, USA.

Bell, J.F., Bramley, T. & Raikes, N. (1998) Investigating A-Level mathematics standards over time *British Journal of Curriculum & Assessment vol 8, no.2 p7-11.*

Bloom, B.S. (Ed) (1956) *Taxonomy of Educational Objectives Book 1: Cognitive Domain*, McKay New York.

Bruner J. , Oliver R. R., Greenfield P. M., Rigney Hornsby J., Kenny H. J., Maccoby M., Modiano N., Mosher F.A., Olson D. R., Potter M.C., Reich L.C., Mackinnon Sonstroem A. (1966) *Studies in Cognitive Growth*, John Wiley USA.

Cheltenham and Gloucester College of Higher Education and the Midland Examining Group for the Joint Council for the GCSE. (1993) *Setting GCSE Geography questions which differentiate effectively.*

Cresswell, M.J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches. Chapter 5 in Goldstein and Lewis (Eds) *Assessment: Problems, Developments and Statistical Issues*, Wiley.

Dall'alba, G. & Edwards, J. (1981) *The scale of cognitive demand: An instrument for analysing cognitive demand in secondary science.* Educational research and Development Unit. Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia.

de Bono E (1976) *Teaching Thinking.* Maurice Temple Smith, Great Britain.

Edwards, J. & Dall'Alba, G. (1981) Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education, 11,158-170.*

Ericsson, K. A. & Simon, H. A. (1984) *Protocal Analysis Verbal Reports as Data.* MIT Press, Cambridge Massachussetts.

Fisher, B., Russell, T. & McSweeny, P. (1991) Using personal constructs for course evaluation. *Journal of Further and Higher Education*, 15 (1) 44-57.

Fisher-Hoch, H., Hughes S., Bramley T., & Pollitt A. (1997) What makes examination questions difficult? Outcomes of manipulating difficulty of GCSE questions. *Presented at British Educational Research Association Conference*, University of York, September 1997.

Gagné, R. M. (1970) *Conditions of Learning* 2nd Edition. Holt, Rinehart and Winston, USA.

Good, F. J. & Creswell, J. M. (1988) Grade Awarding Judgements in Differentiated Examinations. *British Educational Research Journal* 14(3) 263-281.

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

Griffiths, H.B. & McLone, R.R. (1984) A critical analysis of university examinations in mathematics. Educational studies in mathematics, 15, pp 291-311.

Hughes S., Fisher-Hoch, H., Pollitt, A., Ahmed, A. & Bramley T. (1998) The development of a tool for gauging the demands of GCSE and AL exam questions. Paper to be presented at *British Educational Research Association Conference*, The Queen's University, Belfast, August 1998.

Hughes, S. & Fisher-Hoch, H. (1997) Valid and invalid soures of difficulty in math exam questions. paper presented at 23rd International Association for Educational Assessment Conference, Durban, South Africa. 12 June.

Joint Council for the GCSE (1993) *Setting GCSE geography papers which differentiate effectively.*

Kelly, G. (1955) *The psychology of personal constructs.* New York, Norton.

Kremer-Hayson, L. (1991) Personal Constructs of elementary school principals in relation to teachers. *Research in Education*, 43 15-21.

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

Midland Examining Group for the Joint Council for the GCSE (1992) *Setting GCSE Science papers which differentiate effectively.*

Murphy, R. J. L. (1977) The effect of examiner adjustments on the results of the 1976 re-marking investigation. *AEB Research Report.* RAC/37.

Murphy, R. J. L. (1982) A further report into the reliability of marking of GCE examinations. *British Journal of Educational Psychology,* 52, pp 58-63.

Newell, A. & Simon, H.A. (1972) *Human problem solving.* Englewood Cliffs, N.J.: Prentice-Hall.

Newton, P. E. (1996) The Reliability of Marking General Certificate of Secondary Education Scripts: mathematics and English. *British Educational Research Journal* 22, (4) pp 405-420.

Northern Examining Association for the Joint Council for the GCSE (1992) *Setting GCSE Mathematics papers which differentiate effectively.*

Northern Examining Association for the Joint Council for the GCSE (1992) *Setting GCSE Mathematics papers which differentiate effectively.*

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

OCEAC (1996) *GCE A Level and AS Level History.* Syllabus 9020. UCLES, 1 Hills Road, Cambridge CB1 2EU, 1996.

Parsons J.M., Graham N. & Honess T., (1983) 'A teachers's implicit model of how children learn. *British Educational Research Journal* 9(1) 91-101.

Pollitt, A., Hutchinson, C., Entwhistle, N. & Luca, C. (1985) *What makes examination questions diffuclt? An analysis of 'O' grade questions and answers.* Edinburgh: Scottish Academic Press.

Rigney, J. W. (1978) Learning Strategies: A Theoretical Perspective. In O'Neill HF Jr (Ed) *Learning Strategies.* New York: Academic Press.

SCAA (1992) *Setting GCSE Science Papers which Differentiate Effectively.* A study organised by the Midland Examining Group on behalf of the Inter-Group Research Committee for theGCSE.

SRAC (1990) *A Study of the Demands Made by the Two Approaches to 'Double Mathematics'.* A published investigation conducted by the University of Cambridge Local Examinations Syndicate on behalf of the Standing Research Advisory Committee of the GCE Examining Boards.

Taba, H. (1962) *Curriculum Development: Theory and Practice.* Harcourt Brace & World USA.

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

Tulving, E. (1983) *Elements of Episodic Memory.* Oxford, OUP.

Wilmut, J. (1979) Some aspects of the analysis of structured questions. AEB *Research Reports* RAC/110.