

WHAT MAKES MATHEMATICS EXAM QUESTIONS DIFFICULT?

BERA 1996

Hannah Fisher-Hoch, Sarah Hughes

Research and Evaluation, University of Cambridge Local Examinations Syndicate,
Hills Road, Cambridge CB1 2EU

Abstract

Research is underway addressing the question 'What makes mathematics exam questions difficult?'. Statistical analyses identified 'easy' and 'hard' questions in a mathematics GCSE examination. Qualitative analysis of candidates' errors identified the common sources of difficulty in maths questions. Questions proposed to be at different levels of difficulty have been written and will be trialled on school children in Autumn 1996. The outcomes of the research will feed into the examination writing process in three ways: 1. an understanding of sources of difficulty in questions ; 2. the development of guidelines for examiners writing questions and mark schemes and 3. the development of a cross-subject model of the processes involved in answering a question.

The Research Question

This research asks the question 'What Makes Exam Questions Difficult?'. Stenner (1978) stated that 'If you don't know why this question is harder than that one, then you don't know what you're measuring'. His statement demonstrates a concern for construct validity in examination questions (i.e. that a question measures what it claims to measure). Research into the effectiveness of examination questions has traditionally been concerned with the statistical notions of validity and reliability, but neither of these measures are useful unless a task has construct validity. An understanding of the sources of difficulty in exam questions would enable us to develop questions of higher construct validity and effectively target different levels of difficulty.

Dearing (1996) proposed that:

"An examination is only as difficult as the questions and mark schemes from which it is built up".

Here Dearing identified what should be the key focus of research in examination difficulty: The question and associated mark scheme. Currently examiners composing questions are given guidelines, complying with School Curriculum Assessment Authority's (SCAA)

Mandatory Code of Practice for GCSE's. Relating to the difficulty of questions, the guidelines advise that:

- (c) 'The standard of each paper must be suitable for the range of candidates...'
- (d) '...the question paper must discriminate effectively among candidates...and GCSE papers at the highest tier must provide a suitably demanding challenge for the highest grade to be awarded'.
- (j) 'Where choice from optional questions is offered, such optional questions must make comparable demands on candidates.'

p 2

However, there are as yet no guidelines provided suggesting how to determine '*suitable standards*', what constitutes a '*suitably demanding challenge*' or how to measure and ensure '*comparable demands*'.

Kingdon and Stobart (1988) recognised a difficulty with the GCSE:

'...it is extremely difficult for examiners to pitch, or target, questions at a defined section of the ability range. ...the overlap in any two adjacent papers is considerable, so considerable that in some subjects the higher and the easier paper may themselves even overlap.'

(p42)

Many question writers are highly experienced in developing questions and judging difficulty. However, the tacit nature of their knowledge prevents its wider use and transfer. A shared understanding of difficulty would give novice question setters guidelines and make public the notion of difficulty and thus improve the construct validity of examinations.

Literature on School Examinations

The current research was prompted by the work of Pollitt et al. (1985) who identified categories of difficulty: Concept difficulty - the intrinsic difficulty of the concept itself; process difficulty - the difficulty of cognitive operations and demands made on a candidate's cognitive resources; and question difficulty - which may be rooted in the language of the questions, the presentation of questions and the use of mark schemes in rewarding responses. Appendix 1 illustrates the sources of difficulty in maths questions reported by Pollitt et al (1985).

Since then very little research has specifically aimed to identify sources of difficulty in exam questions. Two types of difficulty can be described. *Valid difficulty* has its source in the mathematical requirements of the question, and is intended by the examiner. *Invalid difficulty*, on the other hand, is caused by features of the question which are not mathematical, for example the language or the context of the question and is not intended by the examiner.

Mobley (1987) provided advice to exam setters writing questions. However, these guidelines were not informed by research. To maximise performance Mobely suggested that examiners follow six guidelines: First identify the purpose of the task and discard any material which is not relevant; Second, clarify the main theme(s); Third consider the use of illustrative material to support or expand the text; Fourth, look again at the purpose of the task and at what the students are required to do; Fifth consider the use of 'organisers' which focus attention on the central themes. And finally, consider how best the material can be made attractive to the readers, since its visual appearance is an important aspect of readability. Mobely suggests conventions for presentation (type face and size, use of headings and tables etc.) to aid readability.

Johnstone (1988), considered the cognitive resources used in the question answering process. Johnstone proposed that three factors control a student's ability to interpret questions: first is the students' re-construction of the meaning of a question. Johnston argued that new information must be compatible with the candidate's existing information to be meaningful. This has implications for the construction of the questions and mark schemes. Second are the limitations of working memory. Overloading working memory may result in brief and incomplete answers. An overload may make further demands on the candidate by requiring them to break the question down into sub-goals and chunk information into usable units for use in working memory. Thirdly, the irrelevant noise in working memory (for example superfluous information or context) drowns out the signal. For candidates with smaller working memory irrelevant noise worsens performance.

To summarise, little research has been reported in the area of question difficulty in examinations. However, the most significant piece of research (Pollitt et al 1985) found three sources of difficulty: Concept, Process and Question. Johnstone (1988) found further evidence for Process difficulty. He found that working memory capacity was a key factor in difficulty. Mobely (1987) predicted that readability of questions was the most important factor in question difficulty.

Related Literature

Other related literature alerts us to influences on difficulty of mathematics tasks such as context (Charrer 1989, APU 1990), the language of mathematics (Rothery 1980) and the development of children's understanding of mathematics (Mayer et al 1984, Hart 1984).

Context

The APU (1990) reported that

"Context has been found to affect success rate from a few percentage points up to 20%"

They showed that performance of lower ability candidates was aided by a degree of context, but a very rich context could reduce performance.

Cockcroft (1992) called for a match between curriculum mathematics and skills required in further education, employment and adult life, suggesting that maths should be taught in contexts in which it would be used in adult life. Cockcroft called for contextualisation in the teaching curriculum in order to encourage the transfer of mathematics taught in schools into useful, life skills. Cooper (1992) suggested that assessment in context is not always appropriate. He stated that

"While items appear to be imbedded in 'real-life' contexts the pupil is more likely to succeed if s/he suspends their knowledge of the 'real' and what they know about how to approach the solution of practical mathematics problems."

This suggests that, although it may be more suitable to teach in meaningful contexts, in the assessment of mathematics pupils may not benefit from questions which are forced into so called 'real' scenarios.

Nickson and Green (1996) found that the degree of context in which a mathematical question is set can affect pupils' selection of the correct mathematical operator. It seems to be necessary to identify the degree of contextualisation which is facilitatory for pupils of different abilities.

Language of Mathematics

Rothery (1980) distinguished 3 broad categories of mathematical words:

1. Words which are specific to mathematics and not usually encountered in everyday language (e.g. hypotenuse, coefficient).
2. Words which occur in mathematics and ordinary English, but involve different meanings in these two contexts (e.g. difference, volume).
3. Words which have the same or roughly the same meaning in both contexts (e.g. fewer, between)

The use of these words in questions must be considered carefully, especially with some age groups. Mayer et al. (1984) identified types of knowledge used in solving a mathematics problem. Firstly, they suggested *linguistic and factual knowledge* is employed, this leads the student to construct their interpretation of what is to be done. This is the initial stage in the question response process and is dependent on reading ability. Assessment of mathematics, it has been argued by practising teachers should assess mathematical, not linguistic skills and abilities. Thus the presentation of the question is key to the validity of the task.

Children's Understanding of Mathematics

Hart (1981) wrote questions which aimed to test children's understanding of mathematics and not just repetition of skills. Using outcomes of trials of these questions on 11-16 year olds Hart classified questions into a level of difficulty. A number of conclusions were made about the difficulty of mathematics questions. Firstly three features were prevalent in the questions which children with a lower understanding of maths could answer:

- questions involved only one or two steps to the solution
- questions contained first operations elements (e.g. addition or fractions)
- questions did not contain abstraction or the formulation of strategies.

Secondly, Hart concluded that there was a need to talk to pupils to assess their true understanding of mathematics. Finally, questions which contained mathematical language that was not part of children's vocabulary were found difficult.

Mayer et al (1984) found that students make four classes of error when solving mathematical word problems. These errors relate to knowledge requirements. The table below shows the relationship between error and knowledge type.

| <u>Type of Error</u> | <u>Type of Knowledge</u> |
|---|--------------------------|
| Translation and understanding | Linguistic and factual |
| Understanding and calling upon Relevant knowledge | Schematic |
| Planning | Strategic |
| Execution | Algorithmic |

The literature presented here has thrown up key issues which could effect the difficulty of examination questions. These are:

- The language of the question
- The capacity of working memory
- The level of contextualisation
- Mathematical (technical) language
- The development of mathematical of understanding

This literature has guided the direction of the research and use of research methodologies and techniques of analysis as well as influencing interpretations of results.

Methodology and Results

The syllabus under investigation was 'Schools Mathematics Project' (SMP) 11-16. One of the defining features of SMP is the

"considerable (but not exclusive) emphasis at all levels on the relationship of mathematics to the real world" (MEG 1994 p 3)

The syllabus is structured with papers 1-6 increasing in difficulty targeting different grade ranges. Pupils sit two adjacent papers.

In 1994 almost 150, 000 candidates sat the examination. For our analyses a sample of 600 scripts was taken. The sample was randomised across examiners and schools. Scripts where the candidate had not completed the last question of the paper were excluded from the sample. Only candidates aged 17 years or under on the date of the examination were sampled. One hundred scripts were sampled from each paper, fifty from the top scoring candidates and fifty candidates with low scores. This sample allowed us to compare the errors made by the two groups of candidates.

The research is structured in three phases: First, statistical analysis has identified questions which candidates found difficult; second, qualitative error analysis identified sources of difficulty in the harder questions, this information was been used to re-write exam questions removing the sources of difficulty; and finally experimental trials are planned to measure students performance on manipulated questions.

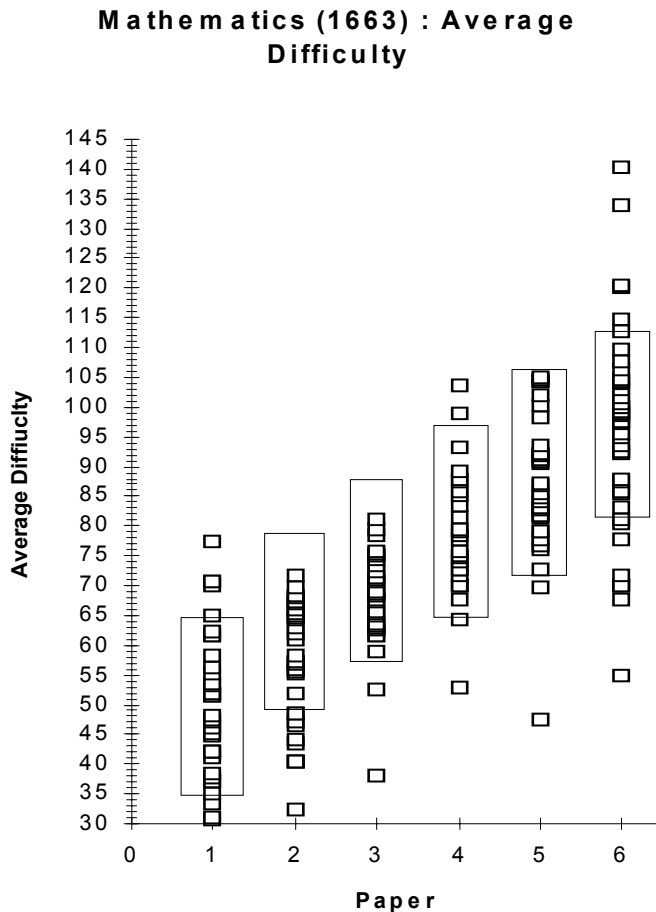
Statistical Analysis

The data from the six mathematics papers were analysed using the Rasch Measurement Model. This is a probabilistic, two parameter, latent trait model which assumes that the probability of success on a question depends upon two variables: the difficulty of the question and the ability of the candidate. These two variables are clearly not the only influences on how difficult an individual finds a question, one could also consider exam nerves, motivation, teaching, etc. These variables vary within the sample and therefore would be difficult to control. The examination process does not take account of these variables (other than in special circumstances) and essentially it is the final score that is recognised, irrespective of the other factors. The Rasch model applies these same assumptions and is thus applicable to this situation.

The analysis provided us with difficulty values for each question and an ability value for each candidate. In order to be able to compare the questions across papers, the papers were equated, using statistics for candidates overlapping papers, and ranked on a common scale. Figure 1 illustrates the difficulty of maths questions across the 6 papers. It shows a general rise in difficulty of questions through paper 1 (targeting grades F and G) to paper 6 (targeting the higher grades D-A*).

The boxes show the range over which each paper was functioning best. The range was identified as that over which there was a linear relationship between the ability of the candidate and the score they achieved. The ranges are indicated in figure 1. Questions outside the ranges were easier or harder than expected.

Figure 1 Difficulty of mathematics questions on 6 tiered examination papers.



Notes:

1. Each square represents a question
2. Boxes show the range over which the test was functioning.

Twenty four questions lying above these ranges (i.e. harder than expected) and towards the top of ranges were targeted for further investigation.

Qualitative Analysis

Candidates' responses were analysed using a qualitative content analysis. The analysis involved two stages:

1. identification of errors made by candidates
2. identification of sources of difficulty (SODs) in the questions.

Appendix 2 shows an example of the outcome of an error analysis. The process of getting from 1-2 required the continual interrogation of the data.

Scripts were sorted into groups of errors and a frequency count showed the most common errors made by candidates.

To identify SODs commonalities in errors were identified through a process of hypothesis making and testing (Strauss 1987). Hypotheses about what feature of the question actually caused the error were made and sources of difficulty were suggested.

Figure 4 Sources of Difficulty in Maths GCSE

| Source of Difficulty | Description |
|---------------------------------|--|
| Command words | The requirements of words which instruct the candidate what is required (e.g. 'explain', 'find', 'estimate', 'state' etc.) were not always the same across questions. |
| Context | The scenario in which the question was set could inhibit the development of a mental model of the question. If the context was inaccessible then the maths was often also inaccessible. |
| Stated principle | If the mathematical topic or concept was not given then candidate had to deduce which topic the question related to. |
| Combination of topics | Difficulty arose in questions which involved more than one mathematical topic. |
| Isolated skill or knowledge | The area of mathematical knowledge or skills required was not well practised by the candidate because it did not overlap with other syllabus areas. |
| Mathematical language | Understanding mathematical terms. |
| Maths v. everyday language | Mathematical and everyday language could have different meanings. |
| Mathematical sequencing | The sequence of the sub-parts of the question did not always follow appropriately. |
| Recall strategy | This was exacerbated when there was a need to recall a strategy that was not given. If the strategy was not recalled then devising a strategy could be more demanding. |
| Alternative strategies | Alternative strategies to those anticipated by examiner could require more steps. This required more of working memory capacity. This had implications for the allocation of marks, where a mark scheme had not anticipated the use of alternative strategies. |
| Abstraction required | Abstract thought was required. |
| Spatial representation required | Spatial skills were required to build a mental model of the question. |
| Paper layout | Physical organisation of the question ordering and or numbering could support or hinder candidates. |
| Ambiguous resources | An unclear diagram affected performance. |
| Irrelevant information | Information appears in question that was not required may have distracted from relevant information. |
| Number of steps | A large number of steps over-loaded working memory and information was likely to be lost. |
| Arithmetic errors | Some questions had more opportunity for making arithmetic errors than others. |

Sources of Difficulty in Maths

The analysis of all difficult questions in maths revealed 19 SODs. They are shown in figure 4. The SODs have impact upon the difficulty of a question at one of the three stages in responding to a question: Reading the Question, Application and Communication.

Figure 5 shows a Model of Question Response developed from a consideration of the SODs. When answering mathematics questions these SODs could affect the performance of candidates at one or more stages in the answering process. The model below shows the first possible stage at which SODs could take effect. The model is hypothesised to be a chronological account of the processes which a candidate experiences and the point at which SODs potentially affect performance.

Figure 5 Model of Question Response and Sources of Difficulty

| Reading the question | Possible SODs | |
|--|--|--|
| Recognise <i>Could I attempt this question? Do I have the knowledge and skills?</i> | Command words Context States principle Combination of topics Isolated skill | Can occur simultaneously with understand ↓ |
| Understand the problem <i>Do I have a model of what the question requires?</i> | Mathematical language Maths v. everyday language Mathematical sequencing | |
| Plan <i>What strategy can I use?</i> <i>What kind of information or data do I need?</i> | Recall strategy Alternative strategies Abstraction required Spatial representation required Stated principle | |
| Application | | |
| Extract <i>Where is the actual data that I need?</i> | Ambiguous Resources Irrelevant information | Loop to execute ↓ |
| Execute <i>I calculate or work out the answer.</i> | Number of steps Arithmetic Errors | ↑ Loop to extract |
| Communication | | |
| Record <i>I write down the solution.</i> | Paper layout | Are there a number of steps? Yes - go to Execute No - Go to check. |
| Check <i>Is my answer reasonable?</i> | | OK? Yes - finish No - Go to a previous stage |

At this stage in the research this model and the sources of difficulty constitute hypotheses to be tested. The hypotheses will be tested in the experimental phase of the project.

Discussion of SODs and Model of Question Response.

A number of the SODs support findings and proposals of other writers described in the literature review. For example Mobley's (1987) notion of readability is reflected in the SODs 'mathematical language' and 'everyday versus mathematical language'. Pollitt's (1985) research concluded with a presentation of 'difficulty variables' thought to cause difficulty in maths exam questions. Some of these difficulty variables are similar to SODs discovered in maths exams in 1994. For example, Pollitt's 'Explaining' is a sub-set of what we have termed 'command words'. 'Degree of familiarity' was described by Pollitt as 'degree to which concept is familiar and central to syllabus', this difficulty variable is similar to the SOD we have called 'isolated topic'. Pollitt's work identified difficulty variables, but did not verify them. The experimental phase of our research will go on to verify the SODs hypothesised to be affecting difficulty of maths exams.

Experimental phase

Questions have been manipulated, hypothesised to be at different levels of difficulty by removing or adding a source of difficulty. Examples of manipulations of question 19 are shown in appendices 3 and 4. The presentation of only one question for this paper presents you with only a light example of the work, one question is likely to illustrate only a few of the sources of difficulty emerging from our analysis of all of the questions. Experimental trials of manipulated and control questions with school pupils will be carried out in the Autumn. One problem in devising manipulations of original questions, has been to control sources of difficulty, as when one thing is changed in a question there is often a knock on effect which varies other aspects of the question. It was our intention to vary only one aspect of the question at a time, however, as you will see from manipulations in appendices 3 and 4 the control of SODs was difficult.

Statistical analyses of performance on the trials will identify which sources of difficulty caused students difficulty. The outcomes will test our hypotheses that the sources of difficulty that we have identified do increase the difficulty of questions. Verbal report data from a sample of candidates in the trials will be used to test the model of question response.

Implications and Applications

The outcomes will be applied to the exam writing process to improve the quality of UCLES's school examinations. The findings of research into difficulty of geography questions has been used at a training day for geography examiners writing questions for future examinations.

The project liaises regularly with an advisory group comprising subject specialists involved in the examination development process. Current concerns for the standards of examinations overtime could be helped by an understanding of what makes a question difficult, and thus help in ensuring consistent standards over time, papers and tiers.

Dearing's (1996) proposal that a focus on individual questions and their related mark schemes is valuable. The recent outcry about the lowering of standards in examinations needs to be addressed. Currently comparisons are made between candidates performance at grades. This can give an overall picture of standards, but this information does not provide advise on how to address the problem. The analysis of difficulty at question level can provide guidelines on firstly what difficulty actually is, and secondly, how we can more confidently assure that standards in questions are comparable, both within and across syllabuses.

References

- Assessment of Performance Unit (APU) (1988) *Mathematical Development; A Review of Monitoring in Mathematics*. 1978 - 1982, Parts 1 and 2.
- Carraher T (1989) Negotiating the results of mathematical computations. *International Journal of Educational Research* 13(6) 637-646.
- Dearing (1996) *Review of Qualifications for 16-19 year olds*. Schools Curriculum and Assessment Authority.
- Hart K M (1981) *Children Understanding Mathematics*. London: John Murray.
- Johnstone A H (1988) *Meaning Beyond Readability*. SEC, Newcombe House, London W11 3JB
- Kingdon M and Stobart G (1988) *GCSE Examined*. The Falmer Press: London
- Mayer, Larkin and Kadane (1984) A cognitive analysis of mathematical problem solving ability. In R J Sternberg (Ed), *Advances in the Psychology of Human Intelligence*. Vol. 2. Hoillsdale, NJ: Erlbaum.
- Mobely M (1987) *Making ourselves clearer: Readability in the GCSE*. SEC, Newcombe House, London W11 3JB
- Pollitt A, Hutchinson C, Entwhistle N and de Luca C (1985) *What makes examination questions difficult?* Scottish Academic Press.
- Rothery (1980) *Children Reading Mathematics*. Worcester: College of Higher Education
- SCAA (1995) *Mandatory Code of Practice for the GCSE*. School Curriculum and Assessment Authority and Curriculum Assessment Authority for Wales. March 1995.
- Stenner (1978) Personal Communication to Alastair Pollitt.

Appendix 1.

Difficulty Variables Identified by Pollitt et al. (1985)

1. Stimulus/Concept Difficulty

| Cross-Subject | Example from maths |
|-----------------------|---|
| Degree of familiarity | Degree to which concept is familiar and central to syllabus |
| Abstractness of mode | Degree to which notation is removed from direct representation of quantity. |
| Abstractness of Idea | |

2. Process Difficulty

| | |
|---|---|
| Explaining | |
| Generalising from data | Mathematical generalisations required; specific (practice) items, types insufficient |
| Selection of data relevant to general theme | Recognition of similarities of new instances to learned e.g. necessary |
| Identification of principles from data | An underlying principle must be derived from specific mathematical examples |
| Applying principle to new data | A given mathematical principle must be applied to an unfamiliar type of example |
| Forming a strategy | Candidate must tailor a strategy from learned principles to solve a problem |
| Composing an answer | |
| Cumulative difficulty | Several mathematical operations required for solution |
| Need for monitoring, logical consistency | Errors in bits of computation lead to answers that should be seen to be unreasonable in context of whole task |

3. Question Difficulty

| | |
|-------------------------------|---|
| Open/closed response | Correct answers may be arrived at in different ways; strategy lies with candidate |
| Leaders, cues, clues | Question does not cue candidate into particular data or strategy |
| Tailoring of resources | Only raw data provided; unclued selection of relevant data required |
| Provision of answer structure | |

APPENDIX 2

Examples of error analyses

Example 1:

Paper 1 question 19 part (b) 'reflex angle' tested candidates' ability to recognise a reflex angle within a shape. It had a difficulty value of 70.17, this is above the range for this paper (35-65).

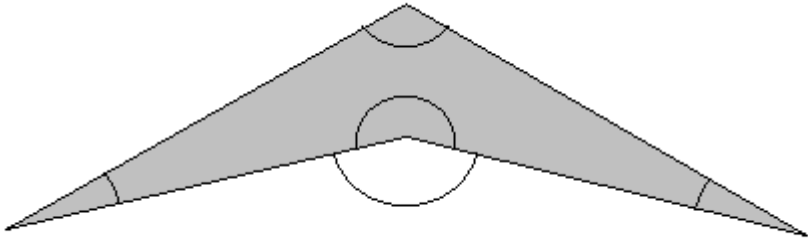
Figure 2 Paper 3 Question 19 (a) - reflex angle and (b) - acute angle (MEG 1994)

19 This shape is called an arrow head.

Mark and label **clearly**

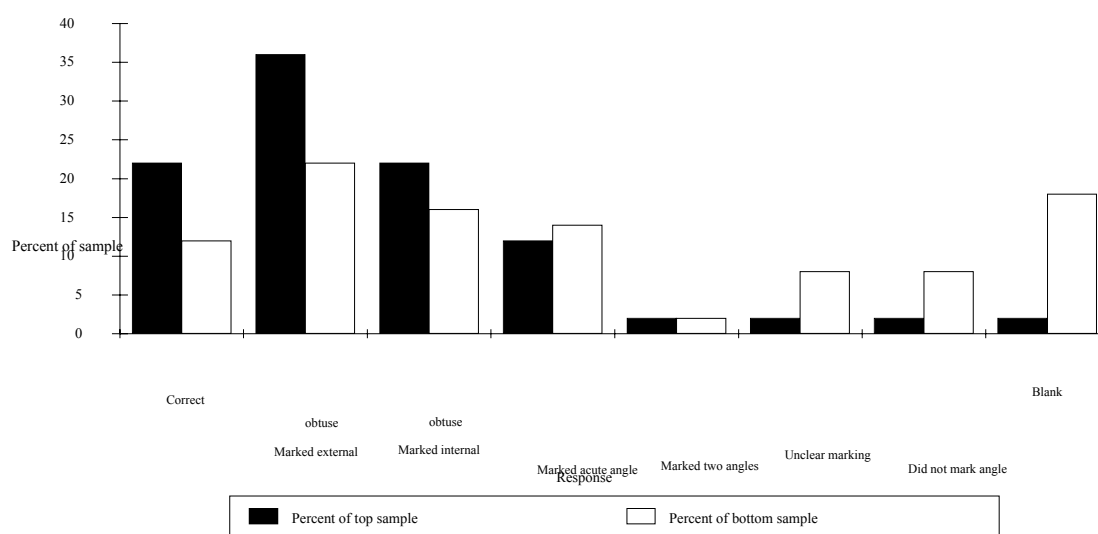
(a) an acute angle, [1]

(b) a reflex angle. [1]



(MEG Mathematics 1663 : Paper 1 : Summer 1994)

Figure 3 Error analysis of 'Reflex angle'.



The identification of errors shows that

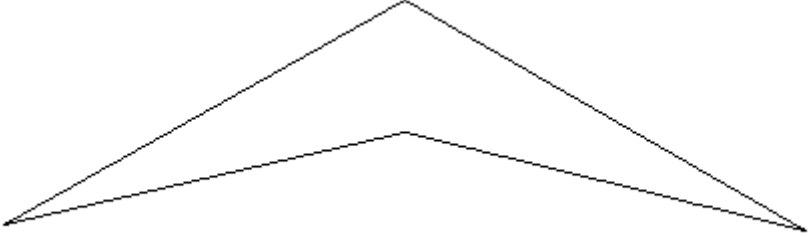
Three sources of difficulty were identified in this question:

1. *The ambiguous resources*: Five angles in the diagram were already marked with an arc. This seemed to cause problems for candidates trying to "mark clearly" already marked angles. At one point on the diagram both the internal and external angles were marked. This caused problems for candidates trying to label one of these two angles and for examiners trying to reward candidates, as it was often unclear which angle candidates were indicating. This was what an 'invalid' source of difficulty - that is an increase in candidate's ability would not increase their chances of getting the question right.
2. *Recall of strategy*: It was necessary that candidates could recall their schema for 'reflex angle'. Candidates who could not recall what they needed to know about reflex angles could not succeed. This source of difficulty was valid.
3. *The context of the question*: The statement 'This shape is called an arrow head' served no purpose. This was an invalid source of difficulty.

APPENDIX 3

Manipulation of Paper 3 Question 19

This question is hypothesised to be easier than the original question because the ambiguity of the angle labels has been removed. Invalid difficulty has been removed.

| |
|--|
| <p>19 This shape is called an arrow head.</p> <p>Mark and label clearly</p> <p>(a) an acute angle, [1]</p> <p>(b) a reflex angle. [1]</p>  |
|--|

APPENDIX 4

Manipulation of Paper 3 Question 19

The irrelevant statement 'This shape is an arrowhead' has been removed. It is hypothesised that this will create a clearer question containing less irrelevant and distracting information.

19 Mark and label **clearly**

(a) an acute angle

[1]

(b) a reflex angle.

[1]

