

Understanding students' minds: the key to writing valid mathematics questions

Alastair Pollitt

University of Cambridge Local Examinations Syndicate

Summary

While technical and statistical techniques are helpful in examination development, the real key to good examining lies in understanding students' minds in the very special circumstances that apply when they are answering examination questions. Two approaches are useful in considering mathematics examination. Our Model Of the Question Answering Process (Pollitt & Ahmed, 1999) describes how students think in the process of attempting questions of any kind, while the psychological literature on problem solving is also pertinent to the special context of solving mathematics problems. By understanding these ideas, and applying them systematically to analysing draft questions, examiners can develop expertise in anticipating how students will behave when they meet new questions. Learning to think like an anxious borderline student is the most useful skill that an examiner can develop.

References:

- Pollitt, A. & Ahmed, A. (1999) *A new model of the question answering process*. International Association for Educational Assessment, Bled, May.
<http://www.uclcs-red.cam.ac.uk/conferencepapers.htm>
- Anderson, J. R. (2000) *Learning and memory*. 2nd edition. New York, John Wiley.

Key words: Psychology, mathematics, examining

Introduction: University entrance maths exams in England

There is almost no use of multiple choice in England's examinations at age 17 or 18. Instead, these exams use multiple mark questions, that is, questions on which the candidate can score 0, or 1, or 2, or whatever up to some maximum mark. A typical example is given in the appendix.

In general, multiple mark questions require the candidate to write an answer, though the mode may involve words, numbers, formulae or drawing, and these answers are then marked by markers specifically employed for the purpose. The marking is usually quite objective, with the markers rigorously following very elaborate marking schemes that specify what is required if each single mark is to be awarded. Considerable time is spent by senior examiners in the preliminary stages, after the examination has taken place, marking a sample of the papers in order to prepare a marking scheme that anticipates most of the kinds of answer that will be seen when the whole population of papers is marked. Further meetings later on deal with any unanticipated responses or other difficulties. The mark scheme for the example is included in the appendix.

This procedure clearly involves much more work for more people than would an examination system that consisted wholly of single mark multiple choice questions, and entails substantial delay between the date of the examination and the announcement of the results. The examination boards are under great pressure each summer to complete the process as fast as possible, since a candidate's entry to a job, training or higher education course often depends on the results. In these circumstances it would be understandable if the boards insisted on using whatever procedures would reduce this delay to minimum. Yet objective testing has not made much impact on British examining. Even in its period of greatest popularity, in the early 1970s, multiple choice was rarely used outside science and mathematics examinations; and almost never, for example, in foreign languages. The principal reason for this failure can readily be seen in these two extracts from a recent handbook for British school teachers:

"However, [multiple mark] questions lose some, if not most, of their value in assessing the higher cognitive skills, the further they move in the direction of multi-choice and other forms of objective testing."

"... questions, such as multi-choice questions, are probably the most valid method of sampling factual recall widely if superficially Extended writing is the most obvious choice for the valid assessment in examinations of higher cognitive skills."

(Lloyd-Jones & Bray, 1986, pp 66,123)

There is a firm conviction amongst British teachers that objective single mark questions cannot be used to assess the 'higher cognitive skills' that really matter, and that these skills can only be assessed by getting the candidates to write their answers.

The second fundamental feature of English examinations in Britain is that the questions are almost never pre-tested before use. The examination system has traditionally been seen as 'belonging' to the teachers (as indeed the quotation above indicates). Most of the marking is carried out by teachers, working in teams supervised by experienced markers. These team leaders, after a few years, may become Principal Examiners and take the responsibility for writing the questions, and for the balance of questions in a particular paper. The PE creates the draft for a paper between 18 and 24 months before it is needed, and submits it to a committee called the Question Paper Evaluation Committee, composed of other senior examiners, and more and more often a language specialist. The aim is to bring as much experience as possible into the writing process, and QPEC members become expert at anticipating how borderline students will respond, how they will understand and misunderstand the task, and what skills will be tested in completing it. The examination board's role at this stage is mainly advisory, and the operation of the QPEC is of great interest to me, as a researcher of the psychology of examining.

A Problem

My first experience of research into mathematics examining for university entrance was in 1990 and concerned the newly introduced AS Level examination. Paper 2 contained 13 questions from which students had to choose 5; on each question students could score up to 12 marks, making a total of 60 marks for the paper. Analysis showed that the questions differed greatly in difficulty, and that there was no correlation between students' ability and their choice of easy or hard questions. Luck in choosing easy ones played a large role in the exam: two students judged to be of equal ability might differ by as much as 20 marks out of 60 or about 2 grades on our A, B, C, D, E, Fail scale for reporting results.

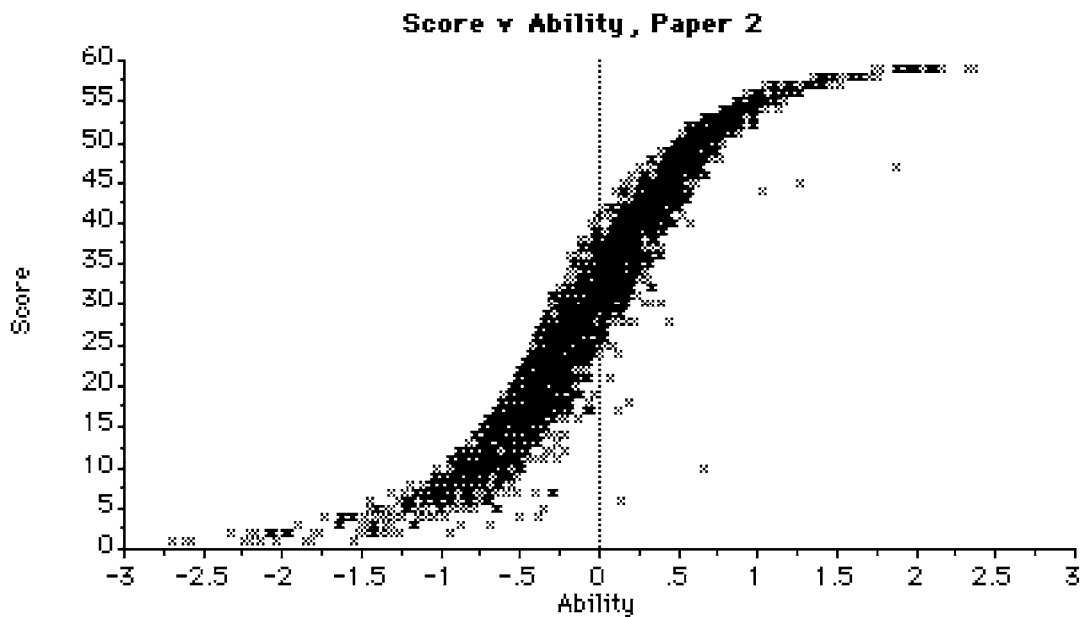


Figure 2

It was clearly unacceptable that luck should be so important in a university entrance exam, and we have now almost abolished question choice in maths exams. I continued, however, to be curious about this effect. We have analysed exam papers in many subjects and have never found differences anywhere else as big as those in mathematics. At the other extreme an analysis of an English Literature exam showed almost no effect of question choice; in fact there was more measurement error arising from differences between different markers than from differences between different questions. Why is this? It seems that Literature examiners are intuitively able to compensate for differences in task difficulty, to be more generous if a particular question is rather difficult - but maths examiners are bound by a more objective marking scheme. In other words, maths questions have a real, intrinsic and fixed level of difficulty. Most of my research since 1990 has addressed the questions "What are the causes of difficulty in exam questions?" and "How can we be sure to include valid and exclude invalid sources of difficulty?"

The problem that remained

Removing choice did not solve all the problems. We noticed that the distributions of raw score for maths questions tended to have very high standard deviations and bimodal distributions, indicating a tendency towards 'all or nothing' scores. The appendix shows the raw score distributions of two 12-mark questions from the 1990 Maths paper, together with a more 'normal' example from Biology.

The consequences of this are serious. Further study showed us that, while internal consistency reliability seems to be quite high for maths exams, test-retest reliability is not: the correlation between students' scores on any two papers in the same exam is surprisingly low. It seems that success on one question tends to lead to success on another, presumably by increasing the student's confidence and motivation. Similarly, one failure tends to lead to more failure. The result is that luck still played a significant role. For example, spotting the correct approach to integrating a particular function – perhaps because you had done a similar example in class – meant that you were well placed to score most or all of the 12 marks in a question, while not spotting it meant you were likely to score very little.

Much more than any other subject, mathematics is a matter of solving problems, and students either succeed or fail on each problem, getting all or none of the marks for that question or question part. The problem could be reduced by making the 'problems' smaller, worth only 1

mark each, but England refuses to go down that road, believing that the sorts of questions we use do represent 'real' mathematical performance.

How then can we create questions of the kind we want that will not be 'all or nothing', questions that will validly assess students' abilities and understanding? This and similar challenges, in maths and also many other subjects, have motivated our research efforts for the last eight years.

MOQAP – the Model of the Question Answering Process

The key to validity is understanding how students think while answering exam questions. Not how we would like them to think or how they 'ought' to think, but what really goes on inside their heads while they are trying to gain exam marks. We choose to define validity in the following way:

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

The purpose of an exam question is to *make* the students' minds do the things we want them to show us they can do. Our business is one of mind control, of helping the students' minds to do the right things, so that they have the best chance of gaining marks. Looking at this from the opposite point of view, we must avoid any feature of a question that will tend to mislead students into doing things we do not want to see them do.

Based mostly on our study of examinations in several subjects at age 16, we have developed a model of the psychological processes that go on inside students' heads. The full Model Of the Question Answering Process (MOQAP) has six phases.

0 Learning		
1	Reading		
2	Searching		
3	Matching		Activation
4	Generating		
5	Writing		

The first phase is **Learning**, which happens (we hope!) before the exam and is what we are trying to measure.

The second phase is **Reading** the question. It is during the Reading phase that many misunderstandings and errors occur, preventing the students from showing us what they can do. Recent research show that each word we read *activates* related concepts in our mind. We then build a *model* from these concepts that represents how we understand the task we have to do.

The next three phases of the question answering model are Searching, Matching and Generating. In **Searching** the activation automatically cascades from the concepts triggered as we read to all the other concepts we associate with them, without any conscious control on our part. **Matching** is the process of identifying, out of the huge number of concepts activated as we search, just those ones that are relevant to this task. As soon as relevant concepts and their relationships are identified our minds **Generate** a rough idea of an answer to the task.

For many purposes we consider these automatic processes 2, 3 and 4 together under the name **Activation**.

The final phase is **Writing** which consists of turning this idea into, usually, a string of words and symbols. Since the idea of an answer does not itself consist of words, symbols or images, but a combination of all of these, students often find great difficulty turning it into an appropriate form.

Often, and especially in exams for younger children, the Activation processes are almost entirely unconscious, and problems occurring there can have profound consequences. Consider these two questions:

- 1 A ski pass costs £4.20 per day.
How much would this cost for 7 days? [1]
- 2 A crate of 12 cans of cola costs £4.20.
How much do 7 crates of cola cost? [1]

Although the arithmetic required is the same in each, multiply £4.20 by 7, the questions were very different. In a sample of 14 year olds, 85% answered Q1 correctly but only 59% got Q2 right. Why was Q2 so hard? Some tried to multiply by 12 instead of 7, but the commonest error was to divide £4.20 by 7 or by 12 instead of multiplying. Having read about a crate of cans the children *expected* to be asked the price of one can. The first sentence activated the memory of ‘questions like this’ so strongly that many students failed to process the second sentence with enough care.

Why does this happen, when pupils are not stupid, and are perfectly good readers?

Schemas

First of all, the example illustrates an important feature of how memory and thinking work. To save the effort of building whole new models for every task we meet, we store in our memory pre-fabricated general models for the tasks or activities that occur repeatedly in our lives. As soon as we recognise some feature of a situation, we activate the schema that seems most closely associated with it and then we *expect* to find all the rest of the schema – whether it is there or not.

In the example, the students activated a schema about finding the price of one can, and this expectation swamped the actual task for many students. At a higher level, students may think they recognise a problem as ‘a Poisson distribution problem’, or ‘an integration by parts’; if they are wrong, they may waste a great deal of time before they see their error.

Stress

The second part of the explanation concerns a well known feature of examinations – they are a very stressful experience. You will all remember the anxiety and worry that exams provoked, especially when the stakes were as high as in a university entrance exam. You will also remember how important the clock was, and how you had to keep checking to see that you had enough time left for the remaining questions. Both of these aspects of stress strongly affect the part of our mind that is called ‘working memory’.

The essential point is that working memory has a limited capacity, that is we can only deal with a small and fixed number of separate ideas at one time – often said to be about seven. Experts (and therefore good students) deal with this by *proceduralisation*, combining several small concepts or rules into a ‘routine’ that can be represented in working memory as a single idea instead of as several. The way that computer programs group commands into ‘procedures’ or ‘subroutines’ is an accurate analogy for this. But everyone, whether expert or not, will find that dealing with the two kinds of examination stress will use up some of their working memory capacity, perhaps reducing it from ‘about 7’ to ‘about 5’ ideas at a time.

I remember an occasion when I could not see the answer to an integration problem even after many, many minutes puzzling at it; when I left the examination room another student said “You had to substitute y for $(x + 2)$ ”. It was simple when he said that, and I felt very foolish. Examination stress makes it more difficult for us to spot unusual aspects of a problem, to monitor our progress towards a solution, and to check our working. Have you had experiences like this?

Modeling mathematics problems

Our model of the question answering process (MOQAP) applies to questions of any kind in any subject. There is, however, one way in which mathematics examinations are, if not unique, then at least rather special, and it is this property that explains the tendency towards ‘all or nothing’ scores on maths questions.

Maths questions are often described as problems that students must solve, rather than as questions that they must answer; I have never heard anyone describe an English Literature question or a Geography question in this way. In most subjects we can think of a student having a rough idea of the answer to a question as soon as it is asked, and spending their exam time improving this initial answer; in Maths, on the other hand, we think of a student as tackling a puzzle, trying various approaches until one works, and then producing the answer quite rapidly when the solution is found. There is a different psychological literature that deals with ‘problem solving’ and which must be considered (alongside MOQAP) if we are to understand students’ mathematical minds.

Psychologists describe two strategies that we follow, sometimes deliberately but often without conscious thought, in solving problems - *difference reduction* and *sub-goaling*.

In the first, we judge how far away we are from our goal and take whatever step seems to bring us closer to it, reducing the difference between where we are and where we want to be. We share this problem solving strategy with almost all other animals, as even the simplest insects will consistently move towards moisture or away from light. In general life this strategy nearly always works well, but it is easy to see how puzzles can be created that violate it. Mazes are an obvious example, where the only way to reach the goal is to increase – temporarily – your distance from it. In a well-known experiment chickens who can see food through a fence will not move away from it even after they have been shown a way round the fence to reach it. Under conditions of stress human beings too will find it difficult to solve problems that violate the difference reduction strategy.

Sub-goaling is a more sophisticated strategy, used only by higher mammals, in which a problem is analysed into parts which can be solved more easily, and is an essential feature of any assessment of advanced mathematics. The analysis is often quite explicit and conscious, and research suggests that the ability to identify sub-goals is a major determinant of students’ success, and of the time they will take to solve problems. For maths examiners an essential skill is to judge when a critical sub-goal is too difficult for students to identify, and to decide how much help to give them. Consider this obvious example:

6 Solve the equation $(\cot\theta + \operatorname{cosec}\theta)^2 = \sec\theta$, for $0 \leq \theta \leq 360$.

Few A Level students would be expected to see that the solution requires the sub-goal of establishing the identity:

$$(\cot\theta + \operatorname{cosec}\theta)^2 \equiv \frac{1 + \cos\theta}{1 - \cos\theta}$$

and it was obvious to examiners that this sub-goal should be set as 6a and followed by the original question as 6b. Note that I am **not** saying that few A Level students could solve this problem without help, but I **am** saying that, especially in conditions of exam stress, we could not be sure that the ones who solved it would be the more able students. The danger of ‘all or nothing’ scoring would be too great.

Examples

Thus there are two approaches useful for understanding mathematical examination questions – MOQAP, which deals mainly with unconscious and often linguistic aspects of the process, and problem solving which is more concerned with explicit strategies for goal reduction. The two come together quite conveniently in Statistics papers, and I will illustrate some of the difficulties

our examiners have dealt with there. The examples come from an **S2** paper, an advanced part of our A Level examination.

Question 1

Q1 Sixty people each make two throws with a fair six-sided die. Calculate the probability that at least 4 of the sixty obtain two sixes. [5]

Calculation of exact binomial probabilities, using a calculator, belongs in **S1**, a lower level paper than **S2**. The intention was that students should use the Poisson approximation to the binomial, with $p=1/36$, but QPEC committee members realised that the question did not specify this, and modern calculators can evaluate $(1-p)^n$ as $\exp(n\ln(1-p))$ just as well as they can evaluate $\exp(-np)$. In order to force the use of the Poisson approximation, the question was amended to:

Q1 Sixty people each make two throws with a fair six-sided die. Using a suitable approximation, calculate the probability that at least four of the sixty obtain two sixes. [5]

Note that '4' was also changed to 'four'. Students will notice when numerals are used rather than words for numbers, will suspect there is some profound significance in the use of numerals, and will worry if they cannot understand it.

One member of the committee asked that the question be "made a little more accessible". It was Question 1 in the paper, and it is good practice to start with an accessible question or two. Note that *accessible* does not necessarily mean *easy*, but it does mean that students should be able to get started with ease rather than being stuck with an impenetrable puzzle. This examiner suggested "maybe two parts with the first part asking for the probability of 2 sixes from 2 throws of a die. Then in part 2 we could hint at a suitable approximation". His first suggestion was not accepted, as the committee felt it was too easy to calculate $1/6 \times 1/6$.

But the examiners were "astonished" to find how many students did in fact use $p = 1/6, 2/6$, or even $1/12$, all of which are not only wrong, but also call into question the appropriateness of the Poisson approximation. Perhaps an explicit statement of the sub-goal would have helped.

Study of the errors made indicates that the problem was primarily linguistic. Many students misread the question, not understanding that the requirement was a *double six* from a single throw of two dice. They seemed to imagine two separate throws of two dice; yet no examiner felt the wording was really ambiguous. The Principal Examiner concluded that he ought to have used the phrase *double six* quite clearly, and at the beginning of the question, as in:

Q1 Sixty people each make two throws with a fair six-sided die. They are trying to throw a double six. Using a suitable approximation, calculate the probability that at least four of the sixty obtain double six. [5]

One of the very general rules of psycholinguistics is that readers pay more attention to the beginning of a sentence than the end, and they start to build their model of the problem using the first things they read. In the version that the students saw, the phrase 'two sixes' appeared only at the very end, and was not read with sufficient care. Many students built a wrong model.

Question 6

8 On average a motorway police force records one car that has run out of petrol every two days.

(a) Using a Poisson distribution, approximated where appropriate, calculate the probability that:

(i) in one randomly chosen day, the police force records exactly two cars that have run out of petrol, [3]

(ii) in one year of 365 days, there are fewer than 205 days on which the police force records no cars that have run out of petrol. [5]

(b) One assumption needed for a Poisson model to be appropriate in part (a) is unlikely to be valid. State what this assumption is, and why it is unlikely to be valid. [2]

This version was submitted to the QPEC, after several drafts to get the wording clear. It is very difficult to be accurate in expressing a real world statistics problem, without making the language so complex that students are lost. Notice that three different distributions are involved, since Poisson is appropriate to model the raw frequency data, but a binomial model needs to be used for the probability in part a(ii), which in turn is approximated by the normal distribution. The real difficulty, intentionally, involved the first two distributions. Despite the very careful wording many students failed to see that they needed to use the cumulative Poisson distribution to calculate the probability of a 'zero day' and then use that in the binomial model; they simply took p to be 0.5. In a good question such an error early on **must** not mean that students cannot continue with later parts. This is a good question, for the correct value of p is 0.6065, and students wrongly using 0.5 will still get a plausible answer for part (ii).

Part (b) proved very good at discriminating between candidates who really understood the models they were using, and those who had simply learned to apply formulae when asked. The intention was simply to test whether students knew that the rate of occurrence is assumed to be constant when the Poisson model is applied; in this case the rate of cars running out of petrol should be constant throughout the year; but it is likely that the rate will change either because the number of cars will vary or because the mix of drivers will vary seasonally. Weak students made comments like "You can't have half a car running out of petrol" or "A car is more likely to run out of petrol on some days than others", both of which show serious misunderstandings of what is happening. And we did, of course, also find some students who told us that the same car, or even police cars, kept running out of petrol every two days! Without the stress of a critical examination, surely they would have been more realistic.

In recent years we have tended to ask our students to write more answers like this, to explain their understanding of mathematics, rather than just to manipulate numbers and symbols.

Lessons

Don't write 'clever' questions, using unconventional methods. These are ideal for teaching, but unfair in testing.

Don't use 'trick' questions that look as if they are one thing, but are actually something else. Such questions bring too much luck into the measurement process.

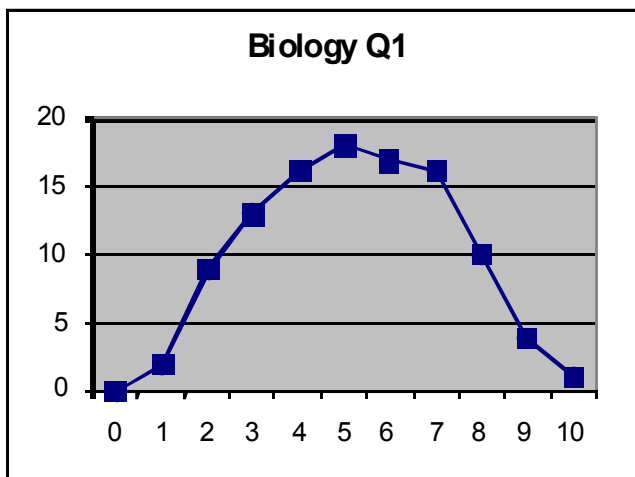
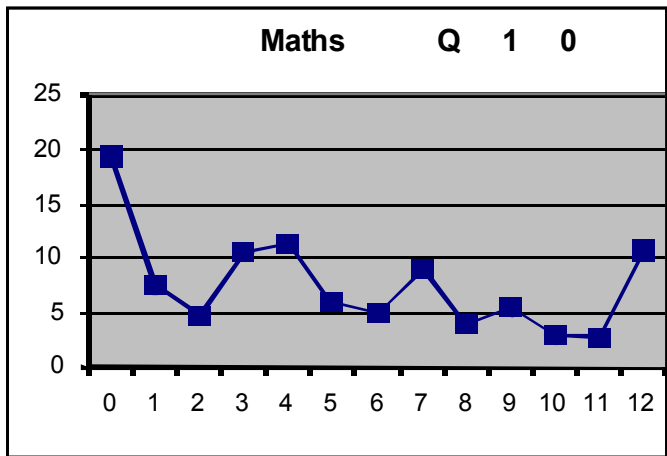
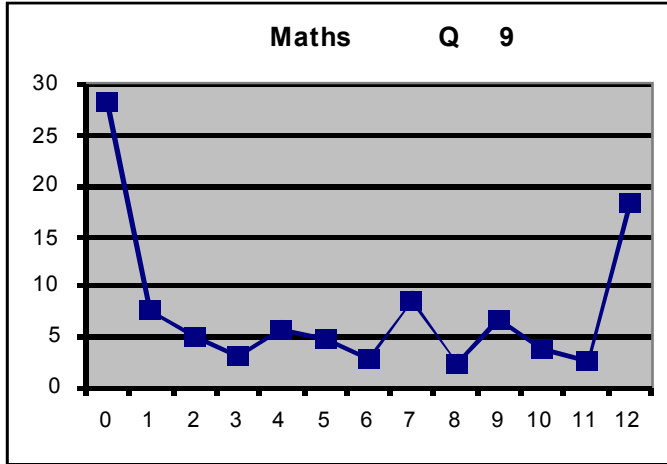
The purpose of a maths examination is to find out how much mathematics each student knows, that is we want to assess Phase 0 of MOQAP. Help students show you their knowledge by avoiding undesirable demands – reading comprehension, bias, or trivial issues like remembering to give the correct units in calculations.

Never think that '*a good student ought to see that . . .*'; remember that stress makes every student vulnerable to any traps and tricks, whether you meant them or not. It is true that expert students will avoid most of the traps, but it is also true that nervous students will tend to fall into them more than confident ones – we are measuring maths, not personality.

Don't let the question get in the way. Express the task clearly, using natural language. Don't put too much complexity in just to ensure that the question is logically flawless, if the result is that many students can't understand it.

Learn to think like a nervous borderline student. Think your way carefully through the MOQAP, trying to anticipate all the possible ways that anxious students might mis-understand, mis-associate, or mis-express their ideas.

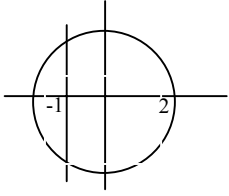
Appendix A: Examples of raw score distributions from A Level exam papers



Appendix B: A typical A Level mathematics question

- 8 The complex number z satisfies the equation $|z| = |z + 2|$. Show that the real part of z is -1 . [2]
- The complex number z also satisfies the equation $|z| = 2$. By sketching two loci in an Argand diagram, find two possible values of the imaginary part of z , and state the two corresponding values of $\arg z$. [5]
- The two possible values of z are denoted by z_1 and z_2 , where $\text{Im } z_1 > \text{Im } z_2$.
- (i) Find a quadratic equation whose roots are z_1 and z_2 , giving your answer in the form $az^2 + bz + c = 0$ where the coefficients a , b and c are real. [2]
- (ii) Determine the square roots of z_1 , giving your answers in the form $x + iy$. [4]

Marking scheme for the example

<p>8 EITHER: Locus $z = z+2$ is a perp bisector Hence $\text{Re } z = -1$</p> <p>OR: $x^2 + y^2 = (x+2)^2 + y^2$ Hence $x = -1$, i.e. $\text{Re } z = -1$</p>	<p>M1 A1</p> <p>M1 A1 2</p>	<p>For recognising linear locus Needs mention of points $z = 0$, $z = -2$, or equivalent</p>
 <p>$1 + y^2 = 2^2$ $\text{Im } z = \pm \sqrt{3}$ $\arg z = \pm(\pi - \tan^{-1}(1/\sqrt{3}))$ $= \pm 2/3\pi$</p>	<p>B1</p> <p>M1 A1 M1 A1 5</p>	<p>Both loci correct</p> <p>Using Pythagoras or equivalent</p> <p>Or equivalent correct method for either case Both correct</p>
<p>(i) $(z + 1 + i\sqrt{3})(z + 1 - i\sqrt{3}) = 0$ $z^2 + 2z + 4 = 0$</p>	<p>M1 A1 2</p>	<p>Form equation and expand LHS; allow any equivalent complete method</p>
<p>(ii) EITHER: $z_1 = 2 \Rightarrow \sqrt{z_1} = \sqrt{2}$ $\arg z_1 = 2/3\pi \Rightarrow \arg(\sqrt{z_1}) = 1/3\pi$ or $-2/3\pi$ $\pm \sqrt{2}(\cos 1/3\pi + i \sin 1/3\pi)$ $\pm 1/2 \sqrt{2}(1 + i\sqrt{3})$</p> <p>OR: If $\sqrt{z_1} = x + iy$ then $-1 = x^2 - y^2$ and $\sqrt{3} = 2xy$ $4x^4 + 4x^2 - 3 = 0$ or $4y^4 + 4y^2 - 3 = 0$ $x^2 = 1/2$ or $y^2 = 3/2$ $z_1 = \pm (\sqrt{1/2} + i\sqrt{3/2})$</p>	<p>B1 B1 ‡ M1 A1</p> <p>B1 ‡ M1 A1 A1 4</p>	<p>For $\sqrt{2}$ For either possibility Convert either case to cartesian form Both correct; allow any equivalent exact $x + iy$ expression</p> <p>B1 Both equations correct Form and solve quadratic in x^2 or y^2 Correct single value for x^2 or y^2 Both correct; allow any equivalent exact $x + iy$ expression</p>

Acknowledgements:

I am grateful to my research colleagues, Ayesha Ahmed, Ezekiel Sweiry and Victoria Crisp for help in developing these ideas, and to Principal Examiner Owen Toller for particularly helpful advice on the S2 paper.