The Support Model for Interactive Assessment

Ayesha Ahmed and Alastair Pollitt
University of Cambridge Local Examinations Syndicate

Paper to be presented at the IAEA Conference, Hong Kong, September 2002.

Contact details
Ayesha Ahmed, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU

ahmed.a@ucles.org.uk

The Support Model for Interactive Assessment

Abstract

The two most common models for assessment involve measuring *how well* students perform on a task (the **quality model**), and *how difficult* a task students can succeed on (the **difficulty model**). By exploiting the interactive potential of computers we may be able to use a third model: measuring *how much help* a student needs to complete a task. We assume that every student can complete it, but some need more support than others. This kind of tailored support will give students a positive experience of assessment, and a learning experience, while allowing us to differentiate them by ability. The computer can offer several kinds of support, such as help with understanding a question, hints on the meanings of key concepts, and examples or analogies. A further type of support has particular importance for test validity: the computer can probe students for a deeper explanation than they have so far given. In subjects like geography or science, markers often would like to ask 'yes, but why?', suspecting that students understand more than they have written. We describe a pilot study in which students were given a high level task as an oral interview with varying types of support. Implications of the **support model** for future modes of assessment are discussed.

Three Models for Assessment

Two models for measurement are dominant in the world of summative educational assessment (Pollitt, 1990). These models correspond to two different conceptualisations of achievement, of what it means to be 'good' at something, and almost every kind of educational test we know seems explicable in terms of one or other of them. When two models are so dominant it is easy to be trapped into assuming that it cannot be possible to think of achievement in any other way; but this would be a mistake, and at least one further conceptualisation has some intriguing potential.

The essential question that we want a final examination to answer is 'how much ability has each student?', where 'ability' is a general term for the amount of learning, skill or achievement the student has acquired, with no necessary presumptions about future ability or aptitude. The most obvious and direct way to measure this is to ask the students to do something and see how well they do it. On the assumption that they all try their best, the ones who do it better are judged to have more ability. By referring to a scaled set of criteria or descriptors, the examiner judges which number on the scale best indicates how well each student has performed. We call this the **quality model** of measurement because central to it is the process of judging the *quality* of a performance (it has also been called the Judging strategy). In practice, the method is limited to the sorts of activities in which we can observe students operating in reasonably standard circumstances, such as speaking in a foreign language, playing the piano, or painting in oils. It is not so appropriate for measuring understanding of theories in subjects such as mathematics, science or geography.

In contrast, to measure understanding we need to adopt an indirect approach, since we cannot 'see' understanding. The usual approach consists of setting a series of tasks that vary in difficulty, often starting with easy ones and moving on to harder ones. What we 'see' is which tasks each student can do successfully. There is an underlying assumption with this approach that the tasks making up an exam can be

ordered from easiest to hardest, and that this order is fairly consistent across various groups of students; indeed if the order varies much from student to student the exam is, by definition, unreliable. It follows that the student who can do more tasks can do more difficult tasks, and we therefore say that this student has higher ability. We call this the **difficulty model** of measurement because central to it is the mapping of students' ability to the difficulties of a graded set of tasks (it has also been called the Counting strategy).

To make the distinction between these two models clearer, consider two sporting paradigms that represent them well: ice dance (free skating) and high jump. Competitive ice dancing uses a pure quality model. All ice rinks are equally flat and roughly the same size, shape and temperature; in other words, the task is pretty much the same in every ice dance performance. The skater is expected to go out and perform in a way that impresses the judges as much as possible. In contrast, a high jump competition is a clear example of the difficulty model, consisting of a series of tasks of ever increasing difficulty which continues until everybody has failed. In the ice dance we focus on judging the *responses*, while in high jump we focus on counting successes on the *tasks*.

In summative assessment, the quality model lends itself to reporting *how well* students do, while the difficulty model is more appropriate for reporting *what* they do, two different and possibly incompatible versions of criterion reference. The technical concern in the quality model is the reliability of the judges, whereas in the difficulty model it is the internal consistency of the questions. In assessing performance it is desirable that examiners are judging the same skills to the same standards, but achieving this is difficult and costly. In assessing understanding we need to know that every question is measuring the same trait or skills, and this also is very hard to guarantee.

Virtually every educational assessment falls into one or the other of these two models or involves a mixture of the two. The only obvious exception is assessment by oral interview. The fundamental difference between oral interview tasks and all other kinds is that every student takes a different exam because the interviewer adapts the task to suit the individual student. This interactiveness is used to target the task difficulty for the student, and in particular to avoid the risk of what we might call 'collapse'. A task collapses when either the student just can't cope with the demands, gets anxious and their level of performance get worse and worse, or when the student fails to understand what the task is and so is unable to get started. In Britain oral interview exams are commonly used with the two extremes of academic achievement – in the Certificate of Achievement exams set for the lowest ability 5% of sixteen year olds and as part of the final exams for university undergraduates and postgraduates – but are absent from the experience of students in between.

The interviewer's role is to adjust the difficulty of the tasks, making them more difficult or easier to suit the student. Yet this is not adaptive testing of the kind now familiar in computerised testing, as the adjustments take place *within* a task, and are based on the student's performance in the *same* task, rather than affecting the selection of the *next* task based on the level of success on the *previous* one.

The problem with the high jump competition is that everyone fails in the end, even the winner. Oral assessment in the CoA is used partly in an attempt to avoid this sense of failure with students who have little experience of success. In the **difficulty model** (the high jump) students are usually all too aware of their failures, of how many questions they could and couldn't do. If we need to measure understanding, when it is not appropriate to measure performance, how can we avoid this constant failure?

In the world outside schools people don't fail, they get help. In the workplace it is not in a company's interest to have their workers failing half of the time. One of the functions of supervisors or managers is to support workers in their work until they are sufficiently competent to complete all their tasks without help. This does not of course mean that workers are not assessed during this time, as their supervisors need to know how much help to give; the essence of the assessment is in fact the observation of how much help is needed.

As a rather different example consider learning to swim. 'Failure' here could mean drowning, an unacceptable outcome. To avoid this risk beginners are given many supports – armbands, a rubber ring, a rope round the waist, and the shallow end of the pool. As they progress, these supports are gradually removed, one by one, until the newly competent swimmer can manage without them. Progress in learning to swim can be measured in how much support the learner still needs.

We propose to develop this as a third model for educational assessment - the **support model**. In any context when we need to assess understanding it is possible, at least in principle, to replace the **difficulty model** with the **support model**, to measure how much help students need to succeed instead of how often they fail. Instead of measuring how high a high jump bar they can clear without help we propose to measure how high a level of support they need to clear any bar, however high it seems at first.

The key to this approach is interactiveness. We need to set students tasks they find too difficult, so difficult that no student would expect to succeed on them without help. Then we can systematically provide help to each student until they produce an adequate response to each task. By appropriately scoring the help we give we will arrive at an assessment of the student's current level of ability.

To develop the **support model** we can start with a context in which something like this already occurs: it is used implicitly in some oral examinations, such as the Geography Certificate of Achievement (Ahmed, Pollitt and Rose 1999).

Oral Assessment

At age 16 in England most students take the GCSE examination in several subjects, often including Geography. The Geography Certificate of Achievement (CoA) is an exam intended for students aged 16 who are not likely to be able to achieve a grade in the Geography GCSE. It consists of a written paper, written coursework, and an oral interview. The interview is designed to last approximately 10 minutes and is on a different set topic each year. The students are given a Resource

Booklet to look at in their lessons in the weeks preceding the exam. During the exam *their own teacher* asks each individual student some 'set' questions about the resources, and the student answers orally. These orals are the first re-introduction of oral methods into public examining in UK schools (other than language exams) for 50 years.

A major issue for this exam is the way in which the teachers ask the questions, and the verbal prompts that they give to the students. The teachers are told that they may be flexible with the way in which they ask the questions, and vary the exact wording to suit their own interviewing style; they are also told that they can ask supplementary questions so that students can achieve something positive. Positive achievement is seen by the examiners as crucially important for CoA candidates, and is sometimes sought even at the expense of reliability. In order to give these students a sense of positive achievement, many of the teachers (but not all) guide students' answers by giving prompts when they answer incorrectly or do not answer at all. The teachers use a variety of approaches when prompting the students. These include requesting further information; rephrasing the question into a more structured form; giving the student extra information; or simply repeating the original question.

Ahmed et al (1999) found that the teachers' comments about their prompting in the Geography CoA revealed the conflict that they felt between assessing the students and helping them to achieve something positive. The first comment below illustrates the teacher's wish to measure the student's achievement without distortion, whereas the second comment shows the opposite sentiment.

> 'Prompting is hard to do without 'helping' the candidate.'
>
> '…even if I had prompted too much, and given them too much help, the most important thing was for them to go out of the room with a good feeling rather than going out feeling they'd failed…'

This third teacher sees the conflict most clearly:

> '…but how far can you go to lead them? How far should you coax them?'

The teachers were not told to take into account the amount of prompting they had given when marking the orals. However, seventy-four percent of those involved in the study said that they had tried to. As one teacher said:

> 'They would get higher marks if they had less help.'

The validity of the CoA would be improved if it were based on a clearer assessment model. There seems to be a mixture of all three models in use: the assessment is based on succeeding or failing at tasks (**difficulty model**), the mark scheme is based on performance level descriptors (**quality model**) and the teachers are helping the students by prompting (**support model**). This has come about because of the conflicting aims of the assessors, but it could be avoided if the mark scheme was based solely on the amount of support given to achieve success. The only model capable of satisfying all of these aims is the **support model**.

The Question Answering Process

Observing teachers carrying out CoA assessments gave us some idea of the kinds of prompts assessors use, and what they are designed to achieve. A more systematic approach to prompting is needed, however, if we are to establish the **support model**.

The aim of prompts as support is to give every student a full chance to exploit the opportunities in a question. Our research has for some time been directed towards discovering the many ways in which students can fail to do this, whether by misunderstanding the task they have been set, by working with inappropriate concepts, or by failing to express themselves clearly. The research is summarised in our model of the question answering process (Pollitt & Ahmed, 1999; in prep).

The full model of the question answering process has six phases. The first is Learning, which happens 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. We define a valid question as one that ensures that 'the students' minds are doing the things we want them to show us they can do'; this clearly cannot happen if the students are unable to understand the question. The next three phases of the question answering model are Searching, Matching and Generating. Searching is the spreading activation of concepts in the mind triggered by the reading of the question. Matching is the identification of relevant concepts, from which an idea of an answer is Generated. The final phase is Writing which conists of turning this idea into, usually, a string of words.

The pilot study

We used our model of the question answering process to inform the design of oral tasks to be assessed using the **support model**. When the student had reached an understanding of a full answer they were asked to turn this into a written response. This formed a pilot study of the use of the **support model**, with the aim of applying these techniques to computerised interactive assessment in the future. In practice it is difficult to distinguish the effects of the three psychological processes of Searching, Matching and Generating. We therefore reduced the model to three phases: Reading, Activation and Writing, which enabled us to select and classify prompts for the oral tasks. First the problems students might have while Reading the questions were considered, and prompts were written to address the resulting misunderstandings. Next we considered Activation, and wrote prompts that cued the students to think about particular concepts when answering the questions. Prompts were also written to trigger activation of particular technical terms, and to encourage students to give deeper explanations of an issue. Finally we designed prompts to facilitate the Writing phase. These encouraged students to formulate a clear response in the manner required by the task. In this way, the sequence of the phases of the model constrained the sequence of the prompts to be used in the tasks.

Methodology

The participants were 11 students, 6 boys and 5 girls, from local comprehensive schools. They were all from Year 10 (age 15).

Each student was interviewed individually by the same interviewer. The interviewer read out a question and asked the student to 'think aloud' about the question. The interviewer interacted with the student using the various prompts listed below, in order to support the student towards a full answer.

After two practice questions, the students were asked two questions from past Geography GCSE papers. The questions were targeted for the top half of the ability range for 16 year-olds (GCSE grades A* - C). The students were chosen from the lower half of the ability range (grades C/D - G), so that the questions were difficult for these students. The question used in the study can be found in Appendix 1. The interviewer's prompting was designed to get the students working at the higher level, so that whatever point they started from they ended up with a full mark answer. This gives the students a sense of achievement as well as allowing us to explore assessment using the **support model**.

When each student had reached a complete answer they were encouraged to write this out or to draw a labelled diagram, and at the end they were asked what they thought of this method of assessment.

The prompts

The prompts that were used in the interviews are classified according to the phases of the model of the question answering process as follows.

Reading Phase

Repeating the question
Re-phrasing the question
Helping them understand the question

These are all concerned with ensuring that the student fully understands the task that they are expected to complete. The three different types of prompt in this category give the student different degrees of help with understanding what they have to do. Simply repeating the question can often be enough to get a student started, and re-phrasing gives them another opportunity to understand the question for themselves. 'Helping them understand the question' involves explaining the meanings of the terms used in the question until they have understood the task. The aim is for all students to be able to gain access to the task, so that they can have a go at answering it.

Activation

Giving a concept
Asking for technical term
Asking for an explanation
Asking for a more specific answer

When a student reads a question the words provoke a mental representation of the task (Johnson-Laird, 1983; Gernsbacher, 1990; Garnham & Oakhill, 1996). As this is formed, many concepts are activated in the student's mind. This is an automatic and unconscious process in which activation spreads through a 'network' of concepts in the brain (Anderson, 1983, Raaijmakers & Shiffrin, 1981). Some concepts will become more highly activated than others, and these will be the concepts the student will use to answer the question; that is, these will come to consciousness as the relevant concepts for the task, and will be the ones that suggest an answer in the student's mind. Some students will identify relevant concepts correctly, whereas others need prompting to think about particular ideas to answer the task. Similarly, for some students the technical terms denoting the concepts will be activated, but others will need to be prompted to use them. Some may have thought of the correct term but not have the confidence to use it.

The prompt to ask for an explanation illustrates a common problem in written papers in subjects such as Geography, Business Studies and Science, in which short essay answers are required. Markers often find that the student's answer does not tell them clearly whether or not the student understands the topic being assessed. When they are asked to 'Explain how…' or 'Explain why …' students often give a first level explanation, or a description, which is not as deep as the explanation the examiners were looking for. Examiners often would like to say to the student 'Yes, but why…?'; if the student has not given a deep enough explanation we don't know if they can't or if they just didn't realise what was expected. Interactive assessment can get round this problem. The oral interviewer, and perhaps in future the computer, can prompt the student to give a deeper explanation, or a more detailed description.

Writing

Asking for a conclusion
Structuring the answer for the task
Writing the answer
Clarification of their answer
Give the answer

Even when students have an idea of an answer in their minds, they can still find it very difficult to turn this idea into a written response. The prompts in this category help them to do this, often by referring them back to the original task. For example, in Question 1 some students gave a full explanation of how rainfall flows to a river when it falls on an urbanised area compared to an area of natural vegetation without concluding that urbanisation would increase river discharge. In some cases students had used the correct concepts in discussing an answer but needed to be reminded of these when turning their ideas into a written response.

Very occasionally, a student could not produce a full answer after a long interaction and in these cases they were given the answer. This is the 'ultimate support'.

Affective

Encouragement

Students are under stress during examinations, whether written or oral. There will be anxiety in both contexts, and a constant concern about monitoring time in written exams. The presence of an interviewer may reduce some of this stress for many students, but could increase anxiety for others. In any case, the effect of stress is to reduce the student's processing capacity and it is likely to lower their level of performance.

In an interactive assessment an interviewer can give feedback on performance, but can also give general encouragement; if the assessment is computerised positive feedback could be given at each step.

Sample transcripts

Example 1

This example illustrates the use of the prompts to encourage the student to explain what he means, and to use technical terms.

'Describe how the sandy beach has been formed. You may draw a diagram as part of your answer'

	Ash: Because soft rocks, it's soft material, erosion's occurring under the beach forcing the soft rock back and then you get deposition there
	A: Right, OK. So that's on this headland isn't it?
	Ash: Yes because it's sandstone
	A: Do you think that some of this sandstone is also …
	Ash: Breaking up, yes
	A: So some of that might be …
	Ash: Yeah, like longshore drift maybe
	A: That's it
	Ash: Breaking up and then deposition in there
Asking for explanation	*A: Yes, exactly, that's exactly how it works. Can you explain a bit about longshore drift?*
	Ash: Yes, longshore drift is when you have materials like sand particles and stuff which the waves come in, in the direction of the wind so if it was from the east would come up the beach and down the beach and up and down, up, down eventually you have deposition there.
Technical term	*A: That's it. Do you know what that up, down of the waves is called?*
	Ash: Swash and backwash
	A: Perfect

Example 2

This example shows the use of prompts to activate the appropriate concepts in the student's mind. Once an idea is given as a trigger the student then sets off on the right train of thought towards an explanation. This also illustrates the prompting of the student to conclude their answer so that the student gives an explanation and then concludes by refering back to the original question.

'Explain how urbanisation affects river discharge'.

	S: Urbanisation um .. build up cities um .. river discharge if a city is built on a river they would need to extract water for all sorts of water supplies like purification for drinking and sewerage and that sort of thing. At the mouth of the river there would be obviously less water because they'd taken some out and the water that does come out maybe a little more polluted because there's rubbish falling in there, being put in there etc
Understand question	*A: Right. Could you explain, could you tell me what you understand by river discharge?*
	S: How much water goes from the river into the sea
Clarificaton	*A: So it's how much water is flowing through the river all the time. So you're saying that the urbanisation, loads of buildings would use up water so there might be less water*
	S: Less
Giving a concept	*A: What about the waste water, do you think that might…*
	S: That could be pumped through separate pipes like sewerage, and they often go to um …. A place on the beach where not a lot of people go or sometimes it's pumped out to sea.
Giving a concept	*A: OK, the sort of ideas I want you to think about now are; think about rainfall and what's going to happen to the rainfall when it lands on urbanised areas.*
	S: When it lands on urbanised areas it's going to be tarmac and surface run off is going to occur, it's going to run to the river and that would increase the river discharge. I've sort of contradicted myself there
Giving a concept	*A: Yes, but now you're on the right track of what this question is getting at. Do you know about lag time?*
	S: No
	A: Lag time is the delay, the amount of time between peak rainfall and rain flowing in the rivers. So do you think that lag time might be affected by urbanisation?
	S: I think it would, I think urbanised areas it would flow quicker, less lag time. It would be easier to run off tarmac than it would off grass
Giving a concept	*A: So when the rainfall is falling on the grass ..*
	S: It sinks into the water table and then runs off
	A: So it's slower. Do you want to have a go at writing an answer to that now based on what we've just discussed?

S: Yes. So ignore the people taking water out of the river?

A: Yes

S: All right, sorry.

A: No, no that's fine. That's why we're doing it

S: 'When a city/town is built on a river, there would be large amounts of tarmac and impermeable materials. When it rains the rain water falls on this impermeable rock and as the rock/materials cannot absorb water it would flow down hill, often towards the river. The lag time, when rain falls on an urbanised area, would be less than it would as if it was on grass. This is because the grass absorbs the water soon after impact and carries it down to the water table, it would run off from there. This would take longer.' Is that OK?

Asking for conclusion

A: Great. The only thing I would say is go back to the question and think so have you sort of .. you've explained all the background to it but have you finally said …?

S: Conclusion .. 'In conclusion, the water discharge of a river would be greater after it has flowed through an urbanised area'

A: Brilliant, you'd get full marks for that

S: Really? Is this like an exam question?

A: It's a higher paper question, yes

S: Cool

Example 3

In this example the student simply needs a repeat of the question and then some encouragement to reach a full answer. She starts by saying she doesn't know, and ends with a perfect answer.

'The map in Fig 2 shows a sandy beach. Describe how it has been formed. You may draw a diagram as part of your answer.'

R: This bit here?

A: Yes

R: The beach bit?

Repeat question

A: Yes. So you want to say how the sandy beach has been formed

R: Oh I don't know how the sandy beach has been formed

Encouragement

A: Have a go. What do you think?

R: Longshore drift?

A: Perfect, see you do know.

R: 'The sandy beach has been created as a result of longshore drift. The waves come into the shore at an angle to the beach. As it comes up to the beach it brings sands and pebbles. The swash pushes the material forward where the backwash forces the material back, which causes the sand and pebbles to move in a zig-zag pattern.'

A: Great.

Students' Comments on Interactive Assessment

The students definitely preferred this method of oral assessment to written exams. However, the issue we are concerned with is not oral examining, but interactive assessment and ultimately computer-based interactive assessment. The students' comments below show the attractions of oral examining and set us a challenge: can we design a computerised procedure that does the same?

> 'Actually, it's quite good because you can think it through quite well because what you tend to do is like write what you think and then you just move on and miss the points like you forget to mention swash and backwash and stuff like that. So it's good, you cover it well.'
>
> 'At the start I didn't know what I was talking about and as I went through and talked about it, it came out more.'
>
> 'It helps if you talk out loud you have the time to understand it a bit more. You find it easier to write the answer.'
>
> 'A lot of people have problems writing it down and it's a lot easier when you discuss it.'

The oral method seemed to help in several ways. Prompting students to review what they had just said led them to repond much more clearly and accurately, giving information that they knew, but would otherwise not have included. More fundamentally, the act of talking helped them to clarify their ideas by increasing the activation of relevant concepts and the suppression of irrelevant ones. Unlike other oral assessments we asked the students here to write their final answer, and they did find it easier to do so after saying it out loud.

Some of these points relate to the oral mode of the activity rather than to its interactiveness, but in a computerised system it will still be possible, using prompts, to help students to answer using all the relevant concepts. Those who do understand the concepts being assessed, but have difficulty producing a full written answer, can be supported by appropriate prompts in an interactive computer-based system.

Discussion and conclusions

When this project started we believed that our psychological model of the question answering process represented with reasonable accuracy the actual thinking that most students would go through while answering a typical British examination question testing the understanding of, and the ability to apply, basic concepts in explaining real world phenomena. We used the model to create a template for supportive prompting, a set of generic prompt types that could be applied in an oral examination context. The interviewer responded to the student's responses with whatever seemed necessary to help them progress towards a satisfactory answer, following only the constraint that prompts should address *reading,* then *activation,* then *writing*. Almost all of the prompts the interviewer gave (the main exception being simple encouragement) fitted well into these categories. We judge that the model of

the answering process succeeded in its task of systematising the process of prompting.

This study used only two questions from a geography examination, chosen as good examples of questions intended to assess understanding. It will take some further work to establish rules for generating the specific prompts that will be needed for all such questions. The use of a general theoretical model which has been tested on many academic subjects maximises the likelihood that the same, or a reasonably similar, set of generic prompts can prove useful for questions of other types and in other subject areas.

It was clear from all of the transcripts that the students 'knew' in some sense more than the traditional written examination format would give them credit for, an idea that Plato developed in *Meno* as a general principle of a theory of knowledge. Given that students must form an understanding of a concept before they will be able to explain it, assessment using the **support model** to elicit explanations should provide a more accurate measure of their understanding than traditional written methods are able to do.

It could be argued that it is important to assess the ability to explain without help. However, assessing explanation is generally used as a proxy for assessing understanding, so it is enough that some questions in a paper are designed to assess the ability to explain. It is unnecessary for all questions to do so as the ability to explain is a general cognitive function that should not get in the way of assessing every bit of understanding in a subject.

As a general point we are led to wonder how much a student's score in a traditional examination is determined by *affective* characteristics such as confidence or willingness to take risks in putting forward uncertain ideas rather than by the student's level of knowledge. By reducing some of the barriers to expressing uncertain knowledge, the **support model** would enable students to show us the full range of their understanding.

The students' comments on this method of assessment were all positive; they valued the feedback and saw the prompts as helpful rather than as telling them things they should have known. In the end they wrote good answers and did feel that they had achieved something worthwhile. The aim expressed by, in particular, the teachers of low performing students entered for the CoA is met in this system.

Can we develop an accurate scoring system for interactive assessment? Suppose that a question has a fully correct answer, which is worth 5 marks. At each stage of the process the student's current response will be given a **mark** which will be an integer between 0 and 5. The most direct way to adjust a student's mark for any help that has been given is to introduce a scaling **fraction**, so that the **score** they get will be their mark multiplied by the fraction. Initially the fraction would be set to 1.0, and a student giving a satisfactory response with no help would score 5 x 1.0, the full 5 marks. Each prompt given would lead to a reduction in the scaling fraction as its current value is multiplied by a value less than 1 representing how much help we think the prompt gives. Multiplying the new value of the scaling fraction by the mark for the student's response, which may have been improved as a result of the prompt,

gives the student's score at this stage. Only if the 'ultimate support' were given would the fraction, and score, fall to 0. (For an alternative approach which seems less flexible to us, see van der Bergh, undated.)

By the end of the process a student will have a string of scores for one question. If students are given the freedom to choose whether or not to ask for help, as in some suggested systems, serious problems will follow. The score will reflect not only the student's ability but also their strategic awareness. There will be an optimum amount of help for each student; if they ask for too little they will get too low a *mark*, if they ask for too much their fraction will be small and they will get too low a *score*. Furthermore, our research has shown that students are often not aware of what support they need. It is one thing to know, another to know that you know it, and still another to be aware of what you don't know.

One distinctive feature of our proposal is that the sequence of prompting is fixed on theoretical grounds. In the next round of piloting we intend to continue to use a human interviewer rather than a computer, but use a more standardised way of turning the generic prompts into specific ones. This will simulate the following phase in which we plan to administer the questions on-screen with automated prompting. In this way we hope to be able to standardise the students' experiences without losing the benefits of interactive assessment. As far as possible, students will follow the same path through the task up to the point at which they produce their satisfactory response. Of course a human interviewer will be able to skip some prompts, when the response so far shows that they are not needed. It is likely that computer-based systems will be able to do the same fairly soon, but we expect early computer versions of our system may continue to offer every prompt in turn. We may avoid the problem of lowering the scaling fraction for an unnecessary prompt, by not including its fraction in the computation when it does not lead to an improvement in the student's response. An alternative might be to award a student the maximum score that they achieved at any point in the sequence.

Computer marking of extended responses is now well established, using a variety of techniques (Landauer & Dumais, 1997; Burstein et al, 1996; Vantage Technologies). Our project will need the ability to mark shorter extended responses than these current programs deal with, and this is the focus of another current UCLES research project. (See Mitchell et al, 2002 for a very recent report on the possibilities in this area.) It is likely that intelligent systems will soon be able to learn from the responses of students in order to improve the nature of the prompts used with each question, and to improve the validity of the scaling factors for each prompt. Computer based assessment necessarily works using a bank of prepared questions and mark schemes. Specific prompts for these questions can currently be written by humans and put into the item bank. Ultimately we hope that a computer would be able to generate by itself specific prompts for each question, using the generic prompts given by the model of the question answering process.

One very attractive feature of this style of assessment is the way that it leads students towards an adequate answer, detecting and correcting misunderstandings and confusions along the way. Stroud (1946) claimed that the time spent taking a test

was probably more efficient *as a learning experience* than any other comparable length of time that students were likely to spend in their classrooms (quoted in Ebel, 1972). Whether he was right or not in the case of traditional multiple choice tests, it seems almost certain that this will be true for a good **support model** based test. Indeed, there is considerable interest in developing assessment of this kind for some on-line teaching schemes.

Summary

This pilot has established that it is possible to design a system to carry out **support model** assessment in a reasonably consistent way. The UCLES model of the question answering process gave us a set of generic prompts, at least generalisable to all questions which aim to assess students' understanding of concepts. With suitable preparation, equivalent to writing a good marking scheme, it seems that these generic prompts could be translated into specific prompts for each question that might then be offered to students by computer, as soon as it becomes possible to judge their responses on-line with sufficient accuracy.

References

Ahmed, A, Pollitt, A & Rose, L (1999). *Should oral assessment be used more often?* Paper presented at the British Educational Research Association conference, University of Sussex, Brighton.

Anderson, J R (1983). *The architecture of cognition*. Cambridge MA: Harvard University Press.

Burstein, J, Kaplan, R, Wolff, S & Liu, C (1996). *Using lexical semantic techniques to classify free-responses*. In Proceedings from the SIGLEX 1996 workshop; Annual meeting of the Association of Computational Linguistics, University of California, Santa Cruz. http://www.ets.org/research/erater.html.

Ebel, R. (1972). *Essentials of Educational Measurement*. Englewood Cliffs: Prentice-Hall, p42.

Garnham, A & Oakhill, J V (1996). The mental models theory of language comprehension. In B K Britton & A C Graesser (Eds) *Models of understanding text*, pp 313-339. Hillsdale, NJ: Lawrence Erlbaum.

Gernsbacher, M A (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum.

Johnson-Laird, P N (1983). *Mental models*. Cambridge: Cambridge University Press.

Landauer, T K & Dumais, S T (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **25**, pp 259-284.

Mitchell, T, Russell, T, Broomhead, P & Aldridge, N (2002). *Towards Robust Computerised Marking of Free-Text Responses*. The 6th International Computer Assisted Assessment (CAA) Conference, Loughborough University.

Pollitt, A & Ahmed, A (2002). in preparation. *Improving the validity of exam questions by understanding how students think*.

Pollitt, A & Ahmed A. (1999). *A New Model of the Question Answering Process*. Paper presented at the International Association for Educational Assessment Conference, Slovenia, May 1999.

Pollitt, A (1990). Giving students a sporting chance: assesment by counting and by judging. In Alderson, C. & North, B. *Language Testing in the 1990s*, pp 46-59. Macmillan: London.

Raaijmakers, J G W & Shiffrin, R M (1981). Search of associative memory. *Psychological Review,* **88**, 93-134.

Stroud, J. B. (1946). *Psychology in Education*. New York: David McKay, p476.

van der Bergh, N.(undated). *AIM documentation.* University of Ghent. http://allserv.rug.ac.be:80/~nvdbergh/aim/docs/

Vantage Technologies Inc: http://www.intellimetric.com.:80/

Appendix 1

Question 1

Explain how urbanisation affects river discharge. [5]

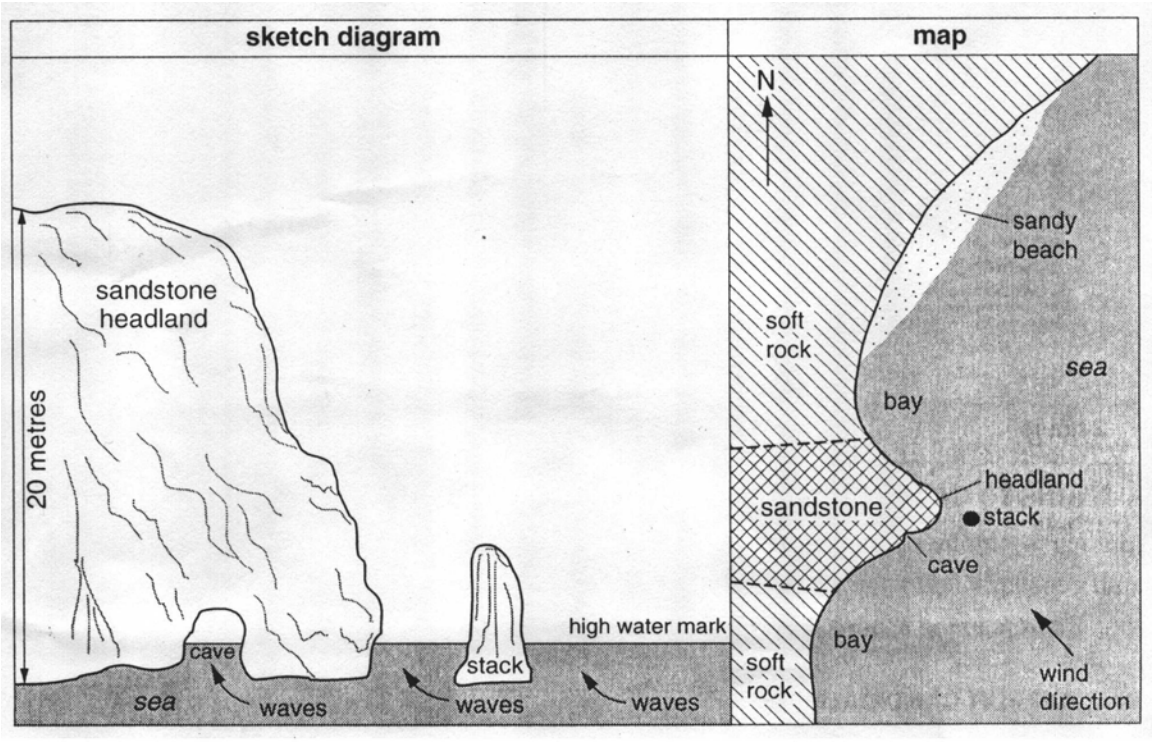
Mark scheme

Impermeable/buildings/concrete/tarmac
water can't infiltrate/soak in
so water runs off surface
water goes quickly to river/shorter lag time
rise in volume/discharge
sloping roofs lead to pipes
disposal of used water
higher density of artificial drainage channels
possible removal of vegetation/fewer trees
water extraction for homes/industry
etc.

Question 2

The map on Fig. 2 shows a sandy beach. Describe how it has been formed. You may draw a diagram as part of your answer. [5]

Fig. 2 A coastal headland



Mark Scheme

Waves are blown in wind direction; waves carry sand

Swash moves material up beach at an angle; backwash moves material back down; longshore drift moves material along coast; movement in direction of current; constructive/spilling waves; push material up beach; etc.

Deposition idea. Material from source/headland/soft rock provides material. Credit labels on appropriate diagram.