12 Computer-based assessment: a platform for better tests?

Hugh Burkhardt
Mathematics Assessment Resource Service (MARS), University of Nottingham, Nottingham, UK, and School of Education, Michigan State University, Michigan, USA

Daniel Pead
Mathematics Assessment Resource Service (MARS), University of Nottingham, Nottingham, UK

Introduction

THE POTENTIAL OF THE COMPUTER as an aid to better assessment has long been thought exciting but has not yet yielded much that is impressive in practice. Once you look beyond simple short items with multiple-choice or other correct–incorrect response modes, there are difficult and well-understood challenges for assessment designers. Nonetheless, the future looks promising. This chapter will explore, and illustrate with some examples, the opportunities and challenges for the computer as a medium for the four key aspects of assessment: task presentation, student working, student response and evaluating student responses. We shall focus on the domain of problem solving in mathematics, science and design technology.

The chapter is illustrated with examples of tasks, mainly from the World Class Arena project. Static text and pictures are not the ideal medium for describing the interactive experience, so a selection of the tasks discussed is available on the internet at www.nottingham.ac.uk/education/MARS/papers/, under the title of this chapter. Some readers may like to work through the chapter online, trying some of the tasks as they arise.

The role of assessment

Formal assessment, whatever its goals, plays several unavoidable roles, notably:

- *Measuring performance against curriculum goals:* this is the traditional goal of assessment.
- *Epitomising the curriculum:* a set of exemplar tests, preferably with mark schemes and examples of student work, communicate clearly the aspects of performance that will be recognised and rewarded. These may or may not cover the declared objectives of the intended curriculum in a balanced way. For this purpose, assessment tasks are clearer than an analytical curriculum description, and much briefer than a textbook.
- *Driving classroom activities:* for 'high stakes' assessment, where the results have significant consequences for students or teachers, the pattern of classroom learning activities of the *implemented curriculum* will closely match the aspects of performance that appear in the test, that is, the *tested curriculum*. 'What You Test Is What You Get' – hence, 'WYTIWYG'.

Traditionally, public high stakes assessment has downplayed the latter two roles. But the attitude that 'We don't assess that but, of course, all good teachers teach it' provokes weary smiles among hard-pressed teachers, and serious distortions of the education of children. High stakes assessment, if it is to be helpful and benign in its effects, must be a balanced measure of what is important, not just what is easily measurable. Any balanced assessment of the goals of most intended curricula implies the assessment of performance on complex tasks involving higher-level strategic skills and substantial chains of reasoning. In this chapter we focus on such tasks.

The role of computer-based assessment

HOW CAN COMPUTERS HELP?

Before we look in more depth at the challenges and responsibilities of assessment designers, we shall describe some examples that show ways in which computer-based tasks can improve assessment.

First, we think it is worth a brief review of some major features of the history of computer-based assessment. Arising from the mindset of programmers, computers have been used to offer intellectual challenges from the earliest days.

- *1950s:* Early computers offered games, puzzles and 'tests'; compilers were designed to identify errors of syntax, and later of style, in computer programs.
- *1960s:* The creators of learning machines, in which assessment always plays a big part, recognised the value of computers for delivering learning programmes. Nearly all these were linear and branch-free, partly because of the 'combinatorial explosion' that follows when one tries to handle the diversity of errors.
- *1970s:* The huge growth of multiple-choice testing in US education enhanced the attractions of automatic marking, in a self-reinforcing cycle.
- *1980s:* A huge variety of educational software was developed to support learning, with less emphasis on assessment. (Ironically, these materials have not had much impact on the implemented curriculum, but are now a rich source of ideas for high-quality assessment that goes beyond the short item).
- *1990s:* Along with the continuing growth of multiple-choice testing, *integrated learning systems*, a more sophisticated development of the learning machines of the 1960s, began to be taken more seriously.

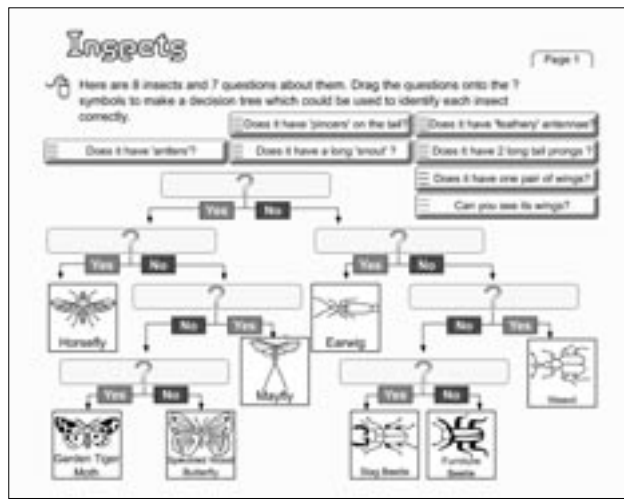
Since the 1990s, the explosive growth of the internet has begun to raise the possibility that *testing online, on-demand* might replace the traditional 'examination day' model, although many technical and educational challenges remain.

In summary, it is well-established that computer-delivered testing can offer:
- *economies* in the delivery of traditional 'paper' tasks;
- *automatic collection* of student responses *if* they can be expressed as simple alphanumeric text, multiple-choice answers or if they provide some form of positional information, as is the case with 'drag-and-drop' responses;
- *automatic marking* of simple student responses that can be mechanically marked without the need for human judgement or interpretation;
- *new types of task presentation* incorporating interactive multimedia elements.

This makes computers valuable for specific kinds of assessment, which are already delivered via multiple-choice or short-answer papers.

Using computers for multiple-choice or short-answer questions

Figure 12.1: *'Insects' task for 9 year olds*



Note: this item, devised by the Mathematics Assessment Resource Service (MARS), 2001 can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

An example of the simple right/wrong approach is 'Insects' (see Figure 12.1, above). This is a classification task, in which students are asked to select appropriate questions for each box on the classification tree from those provided, using drag and drop to input their responses. Essentially a complex multiple-choice task, it suffers from the usual limitations. In this case, it would be more searching to ask the students to *compose* suitable questions. These could be input on the computer but marking them automatically is not straightforward, so there is little or no gain over a paper-based test, and perhaps some loss.

Using computers for more complex questions

The restricted range of task types summarised above should make only a small contribution to any balanced assessment. The major challenges in delivering a wider diversity of assessment via computer include:

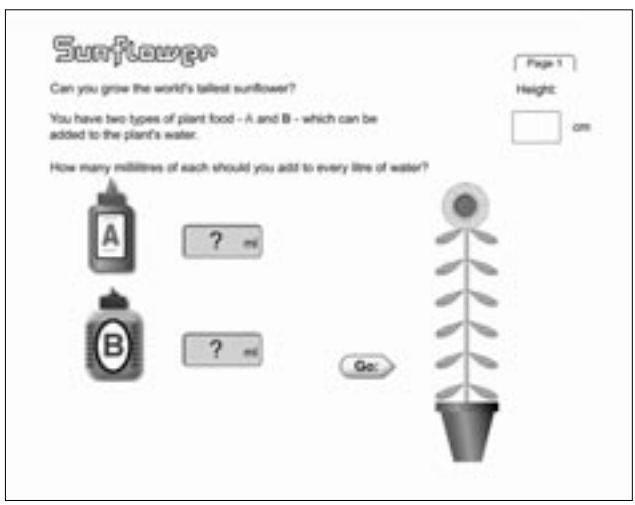
- *exploiting the potential of multimedia* and interactivity and meeting the considerable challenges that these present to the task designer;
- *providing a rich and natural working environment* for the student to work on a complex task;
- *collecting richer, more open forms of response* from the student – without turning every assessment into an ICT skills test;
- *marking richer and more open responses* – methods for marking more complex responses based on artificial intelligence (AI) research have been developed, but they face the long-standing unsolved problem of enabling a reliable and defensible interpretation by computers of open responses in natural languages (to which we shall return).

The remaining examples in this chapter show some of our attempts to create more complex, open-ended tasks, which bring in interactive and multimedia elements. Of course, many groups around the world are contributing to this worldwide effort (see Bennett and Persky, 2002, for a rich example).

'Sunflower' (Figure 12.2, below) is an example of a rich genre, an investigative microworld for the student to explore. In this case, it is a simplified simulation of plant growth. The challenge is to find the amounts of the two nutrients, A and B, which will grow the tallest possible sunflower. The computer accepts number pairs and, with a little graphic support, returns the height that would result. This kind of investigation, in which the computer plays a key role, demands a wide range of important skills. Here the student plays the role of scientist.

The reader may like to consider the aspects of performance on this task that they would wish to capture and reward. We shall return to them later, when we discuss human and computer marking of student responses.

Figure 12.2: *'Sunflower' task for 13 year olds*



Note: this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

Figure 12.3: *'Holidays' task for 9 year olds*

Holidays

To: Advice@holidayhelp.co.wct
From: bozNkat@klubnet.co.wct

Hi – We need a holiday this August.
Must have great night-life not too far away!
We can afford £400 each – but some spare cash for shopping would be nice.
Help!
Boz & Kat

Note: this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

'Holidays' (Figure 12.3, above) is an example of another rich genre – a task that presents a substantial collection of data, gives the students some constraints, and asks them to make inferences and recommendations. This kind of activity, often based on custom-tailored databases, is both important in real life and a common part of the ICT curriculum in many schools for students from the age of eight years upwards. The level of challenge can be adjusted through the complexity of the task: for example, the number of variables, the nature of the constraints and the richness of the response required, can all be modified. This version asks the student to go beyond numerical factors in the selection of a 'best buy' to take into account other qualitative aspects of the client's requests, and to provide a written explanation of their recommendation. Here the student plays the role of expert consultant.

We shall return to these examples later, as well as introducing others to illustrate specific points. First, we shall comment on some general issues of assessment design.

THE RESPONSIBILITIES OF TEST DESIGNERS

Designers of assessment seek to develop tests that enable students '…to show what they *know, understand* and *can do*' across the domain of the assessment (Cockcroft, 1982). A chapter in the previous book in this series (Burkhardt, 2002) discusses design

principles for high-quality balanced assessment that seeks to take this goal seriously; here we shall just summarise some of the key points.

To provide the *opportunity to perform*, any assessment regime should have the properties listed below.

Curriculum balance

Curriculum balance requires that the assessment be fair to all aspects of the intended curriculum. This implies that a teacher who 'teaches to the test' is led to provide a rich and balanced curriculum.

Feedback for teaching and learning

Good feedback is crucial to the effective, self-correcting operation of any dynamic system. In education, a key role of assessment is to provide feedback that is both *formative*, providing guidance for further learning and teaching and *summative*, providing a picture of the students' current level in respect of longer-term goals. In some situations assessment should also be capable of measuring national, and sometimes international, standards, or providing a more detailed diagnostic assessment.

The review by Black and Wiliam (1998b) of research in this area shows the key role that can be played by formative assessment. Selection and accountability tend to dominate discussions of assessment, so that these other constructive roles are often neglected. A better balance would also help counteract the negative view of assessment held by many professionals.

Curriculum value

Curriculum value requires that the assessment tasks should themselves be good learning activities. Tasks such as 'Sunflower' and 'Holidays' both have curriculum value; short assessment items rarely do.

SOME DANGERS IN TEST DESIGN

These responsibilities present great challenges to assessment designers. Much assessment falls far short of meeting, or even of trying to meet, these challenges. Too often, tests consist of rather artificial short items of limited variety. These bear little resemblance to the kinds of task that epitomise the curriculum goals, or which students may meet in real life outside the classroom. This section reviews some of the reasons given for abandoning these goals.

'Good teachers make sure that their students are ready for the test'

The expectation here is that it is the students' responsibility to adapt to the test, and whatever opportunities to perform it may provide – whether or not these cover the learning goals in a balanced way that really allows the students to show what they *know, understand* and *can do*.

'Balanced tests cost too much'

It is true that balanced assessment costs more to manage and to mark than short-item tests with right/wrong answers. However, the true cost of assessment is much more than the fee charged for a test. For high stakes assessment, 'test prep' is a major

curriculum activity in many classrooms – and for understandable reasons, since careers may depend on the results. Teachers we work with in schools often say 'I've got to stop doing mathematics for six weeks now, and get ready for the test.' Test preparation that does not effectively advance learning of the intended curriculum is part of the cost of assessment (Ridgway, 1999). Thus the real cost of an 'inexpensive' test, which may cost just a dollar per student, but leads to six weeks of otherwise relatively unproductive 'test prep', is hundreds of dollars worth of education time. Hence the need for *curriculum value* in assessment tasks, so that 'test prep' is valuable learning.

'These tests are well-correlated with … they take less time and are less expensive'

Reliance on correlation as a justification is as commonplace as it is dangerous. It ignores all but the first role of assessment above – performance measurement (see page 134). Once you consider the curriculum effects or the need for formative feedback, the dangers are obvious.

To avoid such pitfalls, high-quality assessment must be in harmony with the curriculum and its goals. We suggest that, when designing assessment, designers should focus on creating a balanced sample of rich, worthwhile tasks, which cover all the dimensions of the domain, along with a modest proportion of short exercise items on other specific skills and concepts.

FOUR KEY ASPECTS OF DESIGN

What implications does all this have for computer-based assessment? We shall now look in more detail at the four key aspects:

- task presentation: will the students understand the task?
- student working: are the tasks set within a natural working environment?
- student response: do students show what they *know, understand,* and *can do*?
- marking: can we assign proper credit from the evidence we collect?

In the light of this analysis, questions that task designers should ask about each task include:

- Is it a worthwhile task?
- What opportunities does it offer the students to show what they can do?
- Does it need the computer?
- Does it need paper?

To answer these questions in each case needs both analysis and holistic judgement.

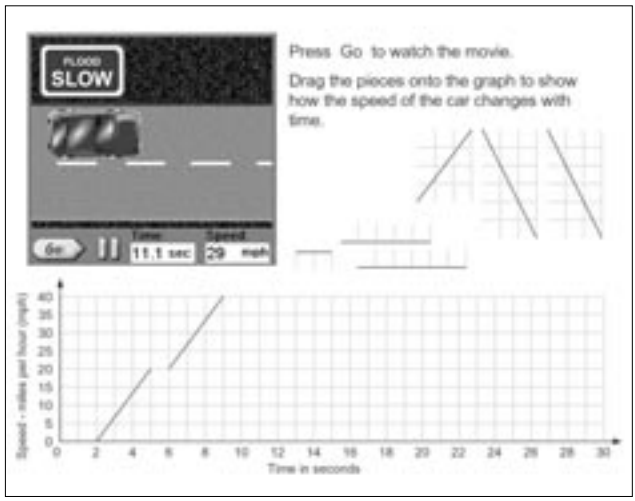
Task presentation

The analysis here is straightforward. Anything that can be presented on paper can be delivered on screen, though one should ask if there is any gain or loss. In addition there are opportunities for:

- *multimedia presentation*, including video and music that can make the problem clearer and more vivid – without the narrowing that a verbal description always introduces. This can be used for relatively short tasks (for example, 'Speed Limit' – see Figure 12.4 on page 140) but seems to have even greater potential for the presentation of

rich open task situations for analysis. Note, however, that there can be practical problems – for example, in a group testing context, headphones are essential to avoid distracting other students.

Figure 12.4: *'Speed Limit' task for 13 year olds*



Note: a newer version of this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

- *rich data* can often be presented on paper but, in assessment as in the real world, custom-tailored computer databases offer opportunities for looking at more data more easily (see 'Holidays', Figure 12.3, page 137).
- *simulations* of practical (or abstract) problem situations for investigation and analysis are a rich and highly promising genre. Examples include 'Sunflower' (see Figure 12.2, page 136) and 'Floaters' (see Figure 12.5, page 141).

One needs, however, to note some negative factors that must be handled by designers:

- *screens hold less information* than a double-page spread on paper, limiting the amount that can be seen at one time. The need to navigate between screens or scroll could cause students to perceive a multi-part task as a series of unrelated items.
- *interactivity can spoil some tasks*: for example, by allowing students to check all their answers, or by encouraging them to persist with trial-and-error searching, rather than think through an analysis. Adapting a conventional task by adding an interactive or multimedia element is liable to cause significant changes in its difficulty and balance.
- *the design and production process becomes far more complex*: facilitating productive interaction between assessment designers, who may have minimal ICT skills, and software developers, who may have no background in education, presents an enormous challenge. Programming aside, the introduction of computers has raised expectations of the standard of graphic design of tests (in World Class Tests this has spilled over to the paper component).

Figure 12.5: *'Floaters' task for 13 year olds*

Note: this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

Student working

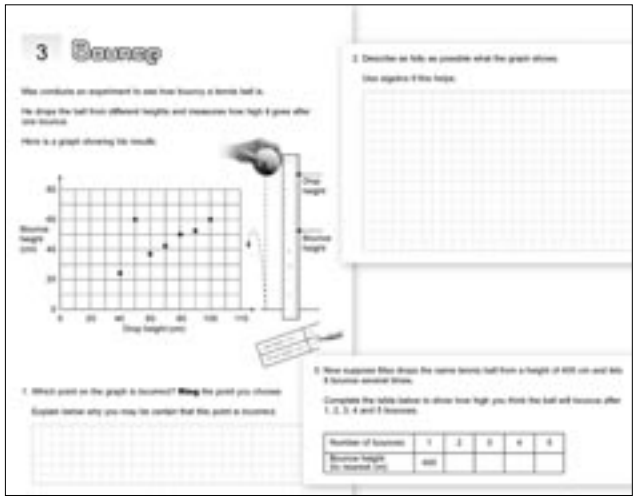
We have stressed the importance, if students are to avoid underperformance, of providing an environment for work on each task that is *natural* to them. For computer-based tasks, this is a more difficult matter, since what is natural varies in time and place, according to hardware and software availability and curriculum (or home) experience.

For example, it is not enough to note that students have some experience of word processing; the issue is whether they are used to tackling a particular kind of task in the environment that the test offers – or whether they should be encouraged to, as a matter of policy. Outside schools, so much work is now done on computers that high stakes assessment could be used to drive curriculum, particularly in school subjects other than ICT. However, as always with change, this approach will only work effectively when such pressure is matched with support and funding for curriculum and professional development that enables teachers and students to respond.

In scientific subjects, for example, rough notes and sketching diagrams, graphs and mathematical expressions, play an important part of working on problems. The computer is not yet a natural environment for such working; it is not even clear that it is a good one, at least for a timed assessment. Office-type tools are biased towards presentation of results for 'publication' rather than their development, while software aimed at working scientists takes time and skill to master.

'Bounce' (see Figure 12.6, page 142) is modeling task, originally presented and tackled on paper. We have also tried it on a spreadsheet with graphing, which seems a good environment for such modeling. Indeed, the output was tidier and more complete *but* the task took an *experienced* spreadsheet user about twice as long to complete as the paper version. A similar effect has been found with other tasks.

Figure 12.6: *'Bounce' task for 13 year olds*



Note: this item was devised by MARS (2001).

There is the further issue of standardisation of software. It is a handicap to be faced with software that is different from that which you regularly use. Even the changes in user interface found between subsequent versions of the same product could pose a distraction on the timescale of a 60-minute test. Possible solutions to this might be:

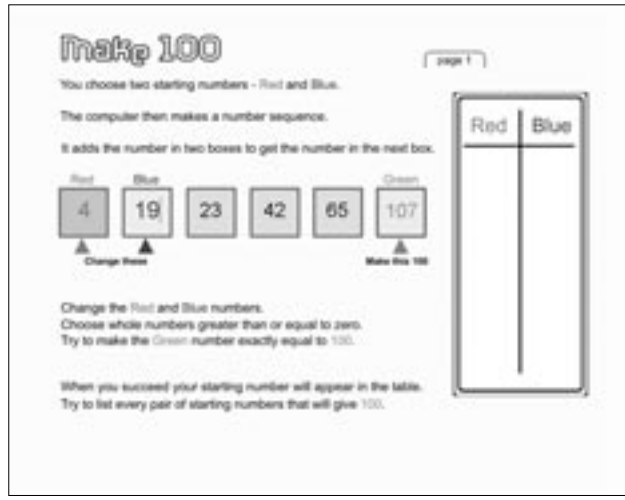
- Impose *universal user interface standards* on software tools embedded in tests. Students could familiarise themselves with these tools before the test.

- Do not embed the software tools in tests, but *allow students to use external applications* with which they are familiar. This still requires standardisation of data file formats, and presents technical challenges in terms of reliability and prevention of cheating (for example, by accessing calculators or communications tools during tests).

- Ensure that the *user interface of the task is simple enough to learn quickly* (in practice, this should take only a few minutes for timed tests). Most current material takes this approach – with the consequent restriction on types of task and responses discussed earlier.

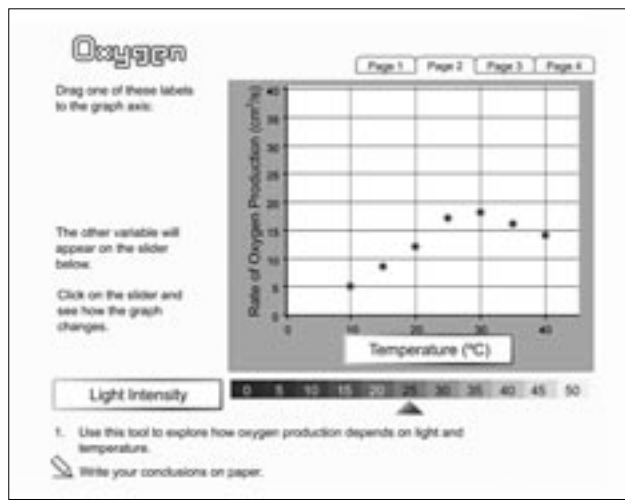
In summary, and despite these complexities, we have found that the computer can provide a natural working environment, at least for:
- *active investigation of simulated microworlds*: for example, 'Sunflower' (see Figure 12.2, page 136) and 'Make 100' (see Figure 12.7, page 143);
- *exploring rich data sets*: for example, 'Holidays' (see Figure 12.3, page 137) and 'Oxygen' (see Figure 12.8, page 143);
- *natural ICT activities*: as discussed above.

Figure 12.7: *'Make 100' task for 13 year olds*



Note: this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

Figure 12.8: *'Oxygen' task for 13 year olds*



Note: this item, devised by MARS (2001), can be tried online at www.nottingham.ac.uk/education/MARS/papers/.

However, there can be problems from:
- time limits on investigation;
- blending computer-based and paper-based work;
- students seeing, and copying, other students' work;
- software familiarity.

Student response

The issues here are closely related to those concerning student working. Indeed it can be argued that, in assessment, the working *is* the response. The following points on the computer's role are important:
- the computer can capture all the interactions of the student with the computer;
- some aspects of the student's work are naturally expressed on the computer;
- the computer provides a natural real-world mode of response to some tasks, though often needing a written response as well.

However:
- only a limited range of the student's thinking is shown through the interactions with the computer;
- it is difficult to capture non-text responses; attempts to do so can spoil a task. For example, in 'Speed Limit' (see Figure 12.4, page 140), though the video presentation of the scene is fine, sketching a graph on paper is a better response mode than dragging the line segments offered onto the graph, which promotes inappropriate multiple-choice thinking;
- the dual mode approach, combining computer-captured and written responses, is often best, but it does not save money – written papers have to be collected, linked to the computer-captured data for that student, and marked by human markers;
- for complex performances, the computer-captured data can be difficult to interpret.

Computers and marking

The core challenge is that, even after 40 years of artificial intelligence, the reliable interpretation by computers of open responses in natural languages is still in general an unsolved problem.

We return to 'Sunflower' (see Figure 12.2, page 136) to give the reader some experience of the kind of challenge that automatic computer-marking represents. For the two nutrients, A and B, the computer captures the successive number-pairs that the student enters, and calculates the resultant height of the sunflower. How far can a computer analysis of only computer-captured responses match human marking with all the information – including, for example, each student's explanation of their approach?

Table 12.1 (see page 145) shows the number pairs tried out by two students working on the 'Sunflower' task; the reader may like to devise an algorithm that will, on the basis of this evidence alone, fairly credit the aspects of performance that we are seeking to assess, namely a systematic search process including, for example:
- controlling the variables, holding one constant;
- trying the combination 0,0 – most plants grow without any added nutrient;
- searching first by orders of magnitude (1, 2, 5, 10, 20 ... not 1, 2, 3, 4 ...);

- going downwards (1, 0.5, 0.2, 0.1, 0.05 ...) as well as upwards;
- systematically homing in on a maximum.

Table 12.2 (see page 136) shows such an algorithm. The reader is invited to judge the extent to which the algorithm is likely to give marks that reflect the elements of performance above. The real test, however, is to compare *computer marks* with *human plus computer marks*, or rank orderings, over a sample of real student responses.

Table 12.1: *Two students' attempts at the 'Sunflower' task (see Figure 12.2)*

	STUDENT 1	
Nutrient A (ml)	Nutrient B (ml)	Height (cm)
30	20	0
15	10	0
18	18	0
10	10	0
5	5	0
50	50	0
25	25	0
500	500	0
250	250	0
150	200	0
150	200	0
150	200	0
130	200	0
130	200	0
130	200	0
130	200	0
STUDENT 2		
10	0	391.3
0	10	0
5	5	0
10	5	0
15	5	0
15	1	0
15	0	374.8
5	0	325.5
20	0	276.1
20	1	0
7	0	361.7
13	0	391.3
10	0	391.3
11	0	394.6
12	0	394.6
11.5	0	395.0
11.75	0	394.9
11.25	0	394.9
11.4	0	395.0

Table 12.2: *Algorithm designed for the automatic marking of the 'Sunflower' task (see Figure 12.2)*

BEST VALUES OF A AND B	INFERENCE	MARK (TOTAL 6)
$11 \leq A \leq 12$	• Has held B constant. • Has tried 0 or <1 for B. • Has searched for maximum using integers.	+1
$11.0 < A < 12.0$	• Has used decimal fractions.	+1
$0 < B < 1$	• Has used decimal fractions less than 1.	+1
$0.3 \leq B \leq 0.4$	• Shows some sort of systematic search for B. • Has held A constant.	+1
$0.30 < B < 0.40$	• Has gone to 2 decimal places.	+1
$A = 11.5, B = 0.36$	• Full marks!	+1

What *has* been achieved in the field of computer marking of complex responses? Progress has been made in some areas, usually when ambition is more limited. Considerable work is currently being undertaken, in trying to extend the domain of the possible. We shall outline a few examples below.

Marking of computer programming exercises

A system called CourseMaster (originally Ceilidh) was developed by Eric Foxley at the University of Nottingham (see www.cs.nott.ac.uk/CourseMaster/). It provides the students with 'instant' detailed feedback on their submitted coursework, whilst enabling staff to monitor the students, auto-mark their work and generate reports about student plagiarism possibilities. To assess the quality of the programming, it uses test data to provide a (limited) check that the program meets its specification, together with a set of standard algorithms that measure the efficiency of programs. The system is used formatively as well as summatively: to improve their marks, students can revise their programs as often as they like before submitting them for formal assessment.

Marking of short-response items

There are products that attempt semantic analysis of short responses, including checking vocabulary and some syntax. See, for example IAT's AutoMark product (www.intelligentassessment.com/AutoMarkFAQ.htm).

Both CourseMaster and Automark exploit the limited domains of the task type – respectively programming languages, and the limited universe of discourse of a short item. Other computer marking is more ambitious.

Marking of essays

Several marking systems are now available, although they make no attempt at semantic analysis. Rather they assess quality through indirect measures, notably standard readability measures such as sentence length and 'rare' word frequencies. The specialised vocabulary for the essay in question may be 'learnt' by the system through 'training' on samples of good work. Two systems widely used in the US are eRater (go to www.ets.org/research and search for RR-01-03) and Intellimetric (www.intellimetric.com).

However, for any system that ignores meaning, a question arises: 'If you know the marking algorithms, can you fake good answers?' The answer seems to be 'Yes, in

principle', though skill is required, and instances seem to be rare in practice. For example, in computer programming, the systematic use of comments that explain the structure is an important feature for enabling program maintenance and later development. Ceilidh checks on this aspect. One student got a good mark, though his (copious) comments consisted entirely of: '/*Blah Blah Blah Blah*/'.

The e-Rater team has conducted a study of this question. They invited a group of experts to write essays that would cheat the system. Some succeeded. The winner just repeated the same (no doubt excellent) paragraph many times (again, see www.ets.org/research and search for RR-01-03).

The other issue facing AI-based marking in high stakes assessment is *defensibility* – can the basis of such results be justified in terms which are accessible to students, parents and potential employers – possibly in the face of appeals over grades?

In view of such concerns and their possible effects, the computer is sometimes used only as a back-up 'second marker'. If the computer mark differs significantly from the human mark, a second human marker is called in. This can still produce important economies in the US context, where there is a tradition of distrust of human marking, and double-marking is therefore common.

FUTURE POSSIBILITIES

We have looked at some of the opportunities and challenges for designers of high-quality computer-based assessment with rich, complex tasks that reflect the major curriculum goals. We finish with a summary of what we regard as the main current opportunities for computer-based assessment to contribute to raising the quality of assessment and, through its influence on the curriculum, of education in schools around the world:

- *simulated microworlds* to be investigated on computer, provide an immensely rich genre in many domains;
- *data-based investigations and modeling* are another rich genre for students of all ages;
- *naturally computer-based tasks*, where the computer is the normal working medium for the student, offer other rich genres, including:
  - spreadsheet-based investigations;
  - text annotation, revision or composition, on word processors;
  - critiquing, modifying and creating designs, on computer-aided design software;
  - computer programming as algorithm design, which reflects a major aspect of modern mathematical thinking patterns;
  - multimedia authoring tasks such as using editing software with video material, either provided or created – see Chapter 9 (Heppell, 2003).

As always, the opportunity for substantial high-quality work is greater when the assessment includes a coursework-portfolio element. The student responses to these tasks will be partly computer-captured and partly written or drawn by hand. The long-mooted shift away from the keyboard towards pen-based user interfaces may help to remove paper from the equation, but does not solve the problems of interpreting the responses. Human marking is likely to predominate, with some computer back-up where this increases efficiency.

In addition there will be some *routine exercises*, which will be entirely computer-handled, including both correct/incorrect response and, increasingly, some AI-based

marking of short answers. The proportion of these needs to be modest (say, 20 per cent) in any assessment that purports to reflect the needs of the modern world and the learning goals of most worthwhile curricula.

For the longer term, better AI-based marking of open student responses to rich, complex tasks remains an important area of work but, after 40 years of AI, don't hold your breath. Other work described in this book may make this seem an unambitious agenda, but beware: the history of assessment is full of neat-but-artificial tests that do not reflect the learning goals in a balanced way, and thus undermine the education that society seeks and needs.