UniServe•Science

UniServe•Science was set up by the Committee for the Advancement of University Teaching (CAUT) in 1994. The Committee established a nationwide network of clearinghouses (modelled on the Computers in Teaching Initiative [CTI] project in Britain) for the collection and dissemination of teaching materials throughout the whole Australian university system. Other members of the network are:

- Engineering at the University of Wollongong;
- Health at the University of Newcastle;
- Law at the Australian National University;
- Humanities & Social Sciences at the Royal Melbourne Institute of Technology; and the
- Coordinating centre at the Australian National University.

UniServe•Science caters for all sciences taught at undergraduate level in the earth, life, and physical sciences. It aims to inform university science teachers of materials in the new technologies that are available to help them in their teaching, and whether those materials may be useful. Its main activities include: finding out what materials are being produced in Australia and overseas, organising reviews of packages by teaching academics, and disseminating all this information, by newsletters, catalogues and electronic means. It also seeks to promote the use of IT in teaching by organising networks of personal contacts, by visiting other institutions and giving seminars, and by organising workshops.

Table of Contents

Students want Feedback: Registrars want Marks

Ian Johnston and Mary Peat
UniServe•Science, University of Sydney
director.science@uniserve.edu.au

Background to this workshop

The ever deepening funding crisis in our universities has led us to re-assess all aspects of our teaching — from the amount of time we spend face-to-face with students to the way we deliver lectures and tutorials. All around Australia, departments are modifying their traditional modes of delivery of laboratory materials to incorporate the use of different media, are reducing the staff time spent in “live” lectures, and are searching for additional ways to reduce the overall budget for teaching in laboratory-based disciplines, whilst trying to maintain the quality of what they do. One of the most important issues we are being forced to consider, perhaps as never before, is how and why we assess our students’ progress.

Obviously we must find out how students have gone at the end of their course. It is important that we are able to certify, to whoever meets them next, that they have mastered what we taught them (even if not as well as we would like). Summative assessment, in the form of end-of-semester, unseen, written examinations, is an integral part of our current system. Registrars want marks. But we also seek to monitor our students’ progress as they go, formatively. We set them homework exercises — it is difficult to do much more, particularly in large first year classes. These are (sometimes) marked and returned with comments (ideally soon after they are handed in: in practice, often much later). By and large, students go along with this system, with all its imperfections, because they need to know how they are going. They want feedback.

Unfortunately, it is a sad fact of current academic life that the marking of homework exercises and examination scripts is a chore which is expensive, time-consuming and (all too often) depressing. When budgetary pressures force us to cut back somewhere, it is often the homework marking that goes first (exams being inelastic). Thus we deprive students of virtually the only feedback we ever give them. Surely if any part of our teaching can be (at least partly) automated, it is assignment and examination marking.

“Objective” testing has, of course, been around for many decades, but the introduction of the computer into mainstream teaching organisation has spawned a number of sophisticated packages capable of providing Computer Based Assessment (CBA) on a large scale. In Australia a few university science departments have been using CBA in their mainstream teaching for some years.

It seemed to us at UniServe•Science that a discussion of the ways in which computers could be used in assessment would be timely. We invited representatives from several disciplines to share their experiences, and we put on our second workshop on February 14, 1997, and devoted it to this subject.

What transpired at the workshop

The workshop was planned to be as hands-on as possible so that everyone attending could gain some experience of what was actually being used in science departments throughout the country. There were to be six sessions. In the end, one of the planned presenters fell sick, and only five presented — Fred Pamula, Department of Biology at Flinders University; Halima Goss, Teaching and Learning Support Services, Queensland University of Technology; Peter Ciszewski, Department of Biophysical Sciences and Electrical Engineering, Swinburne University of Technology; Roger Lewis, department of Physics, University of Wollongong; and Sue Fyfe, Department of Human Biology, Curtin University of Technology.

In the middle of this, we felt that it was important to have a plenary speaker, to remind people of the pedagogical issues involved in this kind of assessment. Professor Royce Sadler, from the Faculty of Education, Griffith University, presented these issues in a clear and concise manner. His talk was singled out in many in their final evaluation forms as the highlight of the workshop.

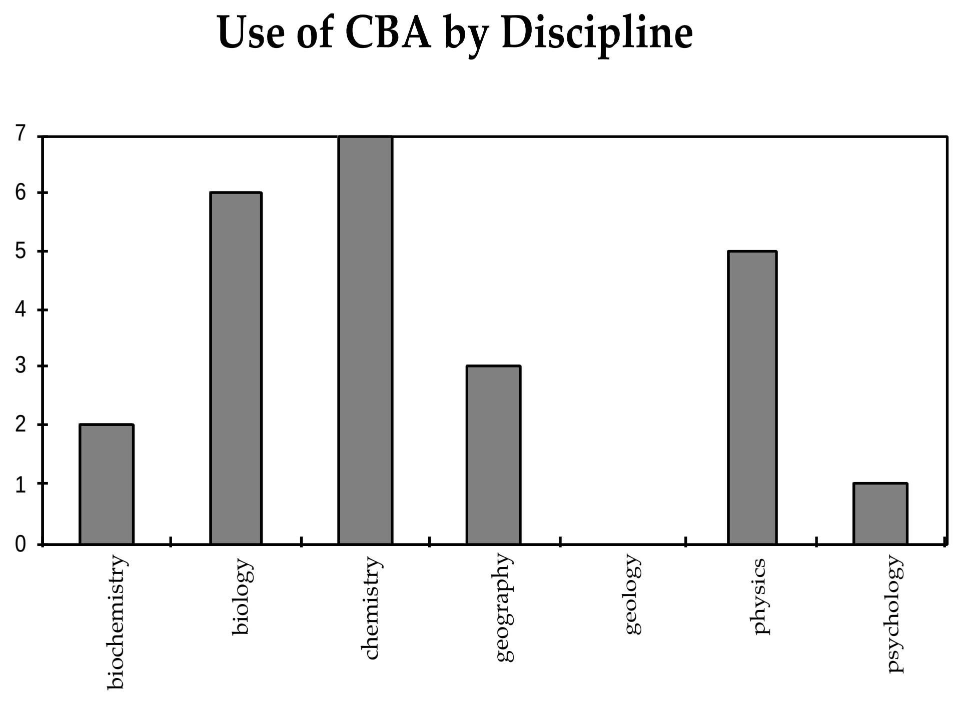
On the day, there were some things that we might have done differently. Since some of those attending had come a long way, it was unfortunate that they could not attend all the sessions. There were a few technical difficulties with some of the software — as there inevitably are in exercises of this kind. There were some problems with the timetable because the weather delayed planes into Sydney by half an hour. But over all we believe the workshop was a success, and what was most appreciated by many of those attending was the chance to get to talk with others doing similar things.

Outcomes of the workshop

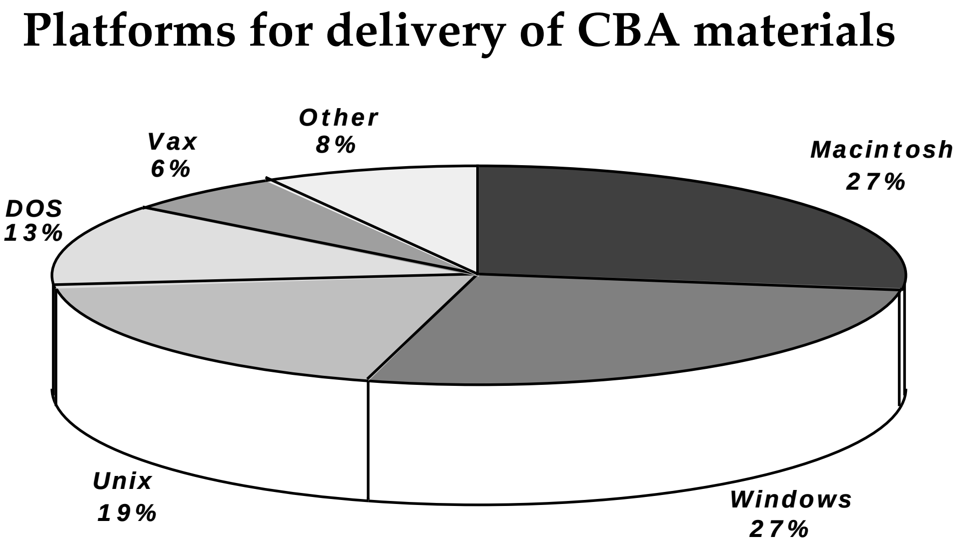
When registering for this workshop, members were asked to fill out a short survey form concerning the use of CBA in their home departments. This provided us with information to start formulating questions that the workshop might address. Then throughout the day, and at the final wrap-up session, these questions were discussed in open forum. The questions that raised most interest were the following.

(1) Degree of usage of CBA.

Of the 67 members attending the workshop, representing about 50 different science departments, 24 stated that they used some form of CBA in their (different) home departments. There were some 170 science departments who were not represented at the workshop, of course, but we have reason to believe, from surveys that we have done in the past, that the number of those who do use CBA but who were not at our workshop is a small number at most. The conclusion then is that CBA is a very small part of the tertiary science teaching scene.
It is interesting to see the distribution of the use of CBA among the different science disciplines. Our survey shows the following:



Though this survey was admittedly very small, the *relative* usage is very similar to what we have found elsewhere[1], viz that Biology and Chemistry are the main users of IT in their first year teaching.
A second interesting statistic that came out of that survey was the proportion of use of different *platforms*.



Those who have the responsibility for deciding what computer hardware to buy could well take note that the split between Macintosh and PC is almost equal.

(2) Feedback vs marks.

Like most things in computer aided teaching, there is a lot of work involved in developing and using CBA. The question which worries us all is: will the students use it? The slogan at the top of this paper suggests that students ought to be willing to avail themselves of what these systems have to offer without compulsion. But one's natural reaction, based on years of teaching experience, is that if you do not make it worth marks, they will not do it.

One cautionary tale that surfaced during the workshop concerned physics students at QUT. Three years ago they were part of the campus wide CBA system that Halima Goss described in her presentation. It was compulsory and the participation rate was essentially 100%. Then a student survey suggested some of the students were unhappy about the inconvenience of the centralised system. The department chose to make it optional. Participation dropped to 10%.

At the wrap-up session at the end of the workshop opinions were expressed that in well designed systems, students have proved that they will work without the carrot of marks. Nevertheless of the 24 departments represented at the workshop which use CBA, 9 of them use it purely summatively and another 12 use it formatively as well as summatively. Only 3 use it purely formatively.

(3) Questions of security.

Opinions were widely divergent on this issue. Some departments adopted the attitude that a lot of thought must be devoted to keeping databases with student marks in them as inviolate as possible, but that it did not matter overly if students found out about some of the questions that were likely to be asked. Others felt that the development of good questions was a time-expensive business (the figure of one-and-a-half hours for a multiple-choice question was quoted by Royce Sadler in his plenary talk) and their usefulness was severely compromised unless security could be guaranteed.

Opinions remained divided, but the need for elaborate security with CBA systems will clearly impact on their cost effectiveness.

[1] See: *UniServe•Science News*, Vol.5 November 1996, p.5

(4) Cost effectiveness.

The biggest question, at least in the view of the bean-counters, is whether CBA can save money. Some institutions (Swinburne University of Technology and the University of Adelaide) claim that they have demonstrated savings in their part-time teaching budgets. Others argued that it is naive to expect that real money will ever be saved, especially when the replacement cost of computers and the time investment in developing questions is properly taken into account; and we should focus our attention on increasing the quality of learning. CBA can relieve teachers of some of the drudge work and leave us free to put effort into what only we, as academics, can do. And then, even if there is no net saving of money or time, the whole exercise will truly be worthwhile.

Conclusions from the workshop

The first UniServe•Science workshop was held nearly one year ago — in April 1996. At the end of that workshop we took stock of what had happened and tried to draw conclusions[2]. It is interesting that the conclusions we draw now are very similar to those we drew then.

This workshop did not come up with definitive answers to any of the big questions (either). But everyone agreed that, while financial concerns may be driving our exploration of matters like CBA, it is still the proper job of the new technology to *enhance* student learning. If it doesn't do that, forget it. However the financial concerns are still there. In this area, as in all others, the cost of using new technology in teaching is frightening. We cannot do much about the burden of continual re-equipment, but (again) we can look to containing the cost of *developing* materials.

Writing questions for use with CBA systems needs a lot of time and energy, and it takes even more time and effort to evaluate whether or not they select out those students who understand our subject as we would like it to be understood. Those departments who have been using CBA for some years have built up item banks of these questions. Those who want to start using CBA are faced with the daunting task of building up their own from scratch. Unless .... Unless we can *share* our item banks with one another.

Like our first workshop, this one brought together many of the active CBA developers and users in the country. The feeling of camaraderie and the spirit of community that it generated were evident, we believe, to everyone who was there. We are very pleased about that. We have something new to think about now. UniServe•Science finishes its initial funding period at the end of 1997, and needs to consider future directions carefully. Perhaps we might capitalise on the goodwill that has been generated and arrange for the establishment of nationwide item banks in which all those working in CBA may share. We will be actively exploring this idea during 1997.

[2] See: "Summary Paper", *Proceedings of Dry Labs Workshop*, (University of Sydney, 1996) p 1.

Assessment Items: Design, Construct, Use, Refine.

Royce Sadler
Faculty of Education,Griffith University
R.Sadler@edn.gu.edu.au

Introduction

The context in which we university teachers are working today is one of many pressures. There are pressures of increased student numbers, with the current concern for mass higher education. There are pressures of reduced funding, which leads to increased and increasing student-staff ratios. There are pressures from the rationalisation of subjects and courses. All these pressures are forcing us to seek out new ways to cope, and preferably to teach better (or at least as well) using the new technologies.

At the same time there are movements towards more flexible teaching and learning, both nationally and internationally. Flexibility is being advocated and implemented with respect to:

- **mode** — There are many more ways of teaching than the lectures, tutorials and laboratories that we all grew up with.
- **location** — Student learning does not have to take place on campus, it can just as easily be done at home, or in other centres.
- **learning rate** —Not all students learn at the same pace. Some learn as they go, some study furiously at the end. There is no reason they cannot all be accommodated.
- **cohort** — It is no longer important for students to learn as part of a larger “class”. If you like, we are moving from “batch processing” to “on-line processing”.
- **enrolment times** — There is no longer need for all students to start classes at the same time of year, nor to abide by arbitrary semester dates.

With all this flexibility entering our teaching, we must learn to be equally flexible with respect to assessment.

Where Computers Come In

What computers can do in this regard is well known, and improving all the time. They can take away many of the routine aspects of teaching, the drill-and-practice, the drudgery for us as university teachers. Furthermore they can do things that we just *cannot* do ourselves. They can construct simulations of complex processes or experiments. They can individualise problem setting for all students. A good example of this occurs in the teaching of statistics. Instead of asking students to analyse a single set of data printed on a handout sheet, individual data sets can be generated from known parameters, one for each student. Then at the end of the exercise class outcomes can be pooled. This is itself another, different learning episode.

We need to design our instructional systems to leave us as university teachers with the interesting tasks, the tasks we as humans can uniquely or best perform with our students. Access to computers is, of course, still an issue, as is equity, but there is no doubt that these are issues that are being solved progressively as the cost of computers gets lower and lower.

However when it comes to considering how the new technologies can contribute to assessment, a definite shift in our thinking is required. Computer assessment is obviously well suited to providing formative assessment for students. It can be administered as regularly as you like, so as to be almost continuous. And it can be used with great flexibility, with tests personalised for individual students. But how should such testing be used summatively? If regular flexible testing is to form a cumulative record of the students’ progress, how do we aggregate their marks?

Currently grading and grading policies are mostly left to individual lecturers. But when students’ marks are returned to central administrators, these marks are usually fitted within approved score ranges, in accordance with expected distributions. Even if there is no explicit policy to this effect marks are implicitly norm-referenced. That is the way that “standards” are “maintained” even if it is not admitted. On the other hand if we shift to flexible assessment, then we will have to adopt a policy of judging outcomes directly against agreed criteria and standards. Students will be judged by their actual achievements. Their marks will not be able to be scaled in relation to those of other students.

These are big changes in our practices and must take place gradually. It will be our job to decide exactly what changes we want to make, and then how to embed them into course structures.

Computer Based Assessment

CBA is popular, largely because it is deemed to be an efficient use of resources and student time, and at the same time offers the hope of an improved standard of interaction with the student. Ideally such assessment can be used interactively to probe the student's understanding to find out the extent of their learning — by sampling domains of knowledge progressively until the critical patch is identified. Having discovered where the student has got to, it is then possible to "slow down" and provide more intensive tuition and testing. Of course, this assumes the existence of a sizeable test bank from which the necessary test questions can be drawn.

The idea of objective testing has an interesting history. What we now call "objective tests" started out as intelligence tests, or puzzles. The main purpose for their use was in *selecting* candidates for further training — in the armed forces and the like. They were constructed with a heavy psychometric influence and the overriding interest was in item and test characteristics — validity, reliability, item analysis, prediction, discrimination and so on — rather than in identifying or promoting learning. Items were trialed, and deleted or retained, modified on the basis of empirical evidence from testing. Educationally this movement had much support from the mass testing industry: school children could be classified, norms could be established, higher education entrance tests (SAT, GSAT, LSAT) could be constructed. The emphasis was all sorting, sifting, ranking, selecting.

In the early and middle 60s matters took a different tack. The idea of behavioural objectives was introduced (Mager). The well-known Bloom taxonomy became popular, as did other taxonomies (e.g. in mathematics). These have proved useful enough, but they all assume a hierarchy of knowledge and thinking skills which is not always valid for all kinds of learning.

Today there are some quite amazing software packages produced to take over this assessment. Computer simulations of real events and processes allows learning and testing of understanding. Computer tutoring and testing combined can be a powerful tool. All this means that the computer provides a new context for developing automated assessment procedures, guided by psychometrics and cognitive learning theories. There are many developments within elementary school education — not counting drill-and-practice — and it has been found possible actually to model the cognitive processes involved in, say, learning arithmetic. Unfortunately it is difficult to extend this kind of research to more advanced education, because the models become much too complex, much too quickly.

Constructing Objective Test Items

The very word "objective" has a nice positive ring about it. That is one reason why objective testing is popular. Many items require a single answer (which is agreed to be correct by experts) or a numerical answer within some tolerance. It is therefore possible to avoid all subjectivity in finding out whether the student knows the answer or not. Examples of this kind of testing are:

- two-state (true/false; fact/inference; yes/no)
- multiple choice (alternatives, options; key plus distracters)
- supply (simple completion)
- matching (premises/options)
- clustered sets of any or all of the above.

Such questions lend themselves to automated delivery (because of the correctness principle). Furthermore it is obviously very easy to give different students different (even unique) tests. However it must be borne in mind that good questions take *time* to construct. A reasonable estimate is one and a half hours for a single multiple-choice item. Ideally any question should be tested by your colleagues before you give it to students. Then it will almost certainly need re-working after students have tried to answer it the first time. All this time adds up.

There are a number of considerations that need to be borne in mind when designing a system to make use of objective testing. Clearly an extensive item bank from which to draw test questions is essential. This requires adequate storage and efficient retrieval. The administration of the tests must be as streamlined as possible: data entry by student directly into the computer does not have to be used, but it is certainly easier when it is. Scoring of the tests should be automated either by computer or mark sense sheet reader. Attention must be paid to the relative difficulty of items in the bank, especially if those items are to be selected randomly for individual student tests. Typically, item banks are organised along psychometric principles. Items within the test bank are evaluated, with information accumulated on how students perform. Within this framework, the usefulness of an item is often judged according to how well it discriminates the good from the weak students, and how reliable the item is in the sense that students of similar ability will give similar responses. And, especially from the students' point of view, the speed at which they can receive feedback is important.

It should be mentioned that some educators harbour reservations about the very concept of this kind of testing. For one thing it forces students to read and think about answers that are wrong. The possibility that inappropriate linkages will be set up in their thinking is very real. Secondly there is the concern that this kind of testing encourages students to work *backwards* to eliminate wrong answers, rather than to work *forwards* to solve the problem for themselves. This skill can become highly developed indeed and the student's ability to pass these tests may not be matched by improvements in problem-solving skills, because the latter are not being practiced.

Constructing Effective Test and Items

A recommended place to start on constructing a test is with what is called in the literature, a **table of specifications**, or a **test blueprint**. Basically, all this is a cross-tabulation of content coverage (that is, subject matter and priorities) against cognitive demands, to achieve the most appropriate balance across subject matter and across cognitive skills (what students are supposed to be able to do with the subject matter). Most untrained item constructors focus on content, not on cognitive processes. For the latter, the literature contains a number of categorisations, depending on the discipline. Here are three examples.

(1) Cognitive demands:
- memory
- comprehension / understanding
- analytic thinking
- evaluation / judgment

(2) Taxonomies or kinds of knowledge:
- factual
- linguistic
- schematic
- strategic
- algorithmic

(3) Findlay's set for mathematics:
- knowledge
- representation
- manipulation
- algorithmic procedure
- attack and structuring

An instructive exercise is to carry out an audit of an assessment program or items, focussing partly on the content, but equally on cognitive requirements, the balances achieved on each dimension and the detailed structure (including wording) of items. The differences between what examiners intended to tap into and what they actually do tap into, is often substantial.

We could, for example, carry out a simple evaluation of a test and its items, using developers, colleagues and consumers. A possible panel of six "judges" might be: two academics who construct items, two independent academics and two students. Ask them for their judgments about items, about the test as a whole, and their (independent) classifications. What is their level of agreement? What do they disagree over?

Effective test items are those for which the language and logic of the item have been carefully scrutinised. The following examples, taken or adapted from Hoffman [1] demonstrate how apparently simple wording can be misleading or capable of interpretations that the setter may not have intended. Think about them first, before perusing Hoffman's comments.

A scientific hypotheses must be

- A. *true*
- B. *capable of being proven true*
- C. *capable of being proven false*
- D. *none of these*

The quickest way to understand the ___________ inevitability of ambiguity in multiple-choice tests is to try making up a multiple-choice item.

- A. *absolute*
- B. *virtual*
- C. *relative*
- D. *inevitable*
- E. *institutional*

The burning of gasoline in an automobile cylinder involves all of the following except

- A. *reduction*
- B. *decomposition*
- C. *an exothermic reaction*
- D. *oxidation*
- E. *conversion of matter to energy*

As a general rule, it is good practice to make the stem of a multiple-choice item into a question which is capable of standing alone. This makes it simpler for the student to process the options. A simple test is to ask yourself: would the stem make sense to a person in the street? If it would, then there is more chance that it will make sense to the student and be unambiguous. There is also more chance that the alternatives will be homogeneous (i.e. about the same thing). Above all, it should always be possible to come up with an answer to the meta-question: What exactly is this item testing?

As for the problem of generating good distracters, it is a sensible idea to have a trial run for your test. Ask the stem as a question requiring a short answer response. Then see what your students give you, and the chances are that you will be able to use these as raw material for plausible alternatives.

If you want items that test beyond recall of information, there are many ways of constructing items to achieve this goal. You might, for example, pose hypotheticals: "Given these constraints, what would be the effect of....?" You might look at extrapolations, taking things beyond the normal ranges: "How would this reaction behave on a planet where .....?". You might test the students' understanding of an argument by stating the initial problem and the desired conclusion, and ask what additional assumption is necessary to reach that conclusion.

Qualitative Judgement in Objective testing

In any university course, the assessment program should be laid down at the planning stage. That means deciding beforehand about what sort of tests will be given — term papers, classroom seminars, written tests, performance tests. The format of each task needs to be thought about, and the marks allocated to each. At each stage of the process professional judgment is required by the teacher.

If objective testing of the kind we are discussing here is to be part of this process then similar judgments have to be made. A few of the questions that require this kind of judgment are the following:

- What marking policies are to be adopted? Do you worry, for example, about students' guessing the correct answers?
- How are the different components of the assessment program to be weighted with respect to one another?
- How do you ensure that you have a sensible sampling of content and skills to be assessed?
- What guidelines do you adopt for the construction of items, the way the stems are to be worded, how distracters are to be chosen?
- Most important of all, what procedures do you lay down so that students can be given adequate feedback?

All these questions must be decided within quite rigid constraints. Teachers have only so much time to devote to this aspect of their teaching. Students also have limited time which must be divided between all that is expected of them. And, of course, the development and maintenance of a workable automated assessment system can be horrendous.

There is a further area of concern to be borne in mind. When designing these kinds of tests it is very easy to adopt an approach driven by learning theories. If you know what are the dominant responses in certain circumstances, it is tempting to develop a testing algorithm based on that knowledge. However, although the responses may be frequent or correlate strongly *in general*, they may not do so *for any one specific student.* Hence generalised feedback routines may actually be dysfunctional (e.g. confusing) to the students who do not conform to the expected pattern. In the worst cases, you may end up with completely the wrong ideas embedded in the student's mind. As yet there is no hard evidence that this does happen but it is as well to be aware of the possibility.

A second area of concern is that the breaking up of learning into components and testing each to a high level of proficiency (in relative isolation from other components) may lead to a fragmentation of knowledge. It makes integration of all the pieces more difficult. Specific items need to include progressive development, connection, contextualization, cross-referencing and cross fertilisation.

By the same token step-by-step feedback on the correctness of a particular move may be detrimental to learning. After all, the best way to solve many problems is to converge on the solution. Too much feedback, offered too immediately, may inhibit various levels of "standing back" and viewing the bigger picture. It does not provide experience in solving bigger problems involving strategic thinking. In other words, there is no point in giving the chook one grain of wheat at a time.

So Where are We Now?

It is worth comparing how a well designed adaptive computer tester might compare with an adaptive *human* tester. With the computer, someone (the software designer) has to imagine all possible "moves" the student might make (or identify them from real student responses) and convert them into formal protocols. With a human tester, the next test or challenge move is invented in context. The human tutor (say, one-to-one), has a sense of the history of immediately preceding error types, and can choose alternative analogies or metaphors as they go along, or re-express the problem in semantically equivalent form. By exploratory questioning, the human can focus on the student's misconceptions in a way that it is very difficult for a computer to do.

It is worth finishing on a re-evaluation of our traditional type of testing, that involving *open, short response items*. Students respond in prose (not just a numerical answer to a well-formed problem). How are these to be evaluated? Consider the value of using the assessment as a proving ground for student understanding, and hence learning. For example, pose a problem and let the students respond. Then pass the student responses in front of other students, asking them to judge the adequacy of the responses, and to provide remedial feedback to fellow students. This would teach students to withhold judgment, to be critically aware of misconceptions, and to learn by themselves being teacher. It is the common experience of teachers that they learn themselves through the very process of teaching others.

References

[1] Hoffman, Banesh *The Tyranny of Testing*. New York, Collier, 1964
[2] Bloom, B.S. (ed) *Taxonomy of Educational Objectives: Book 1, Cognitive Domain* (1956). Old but still a classic and with interesting examples for higher cognitive skills.
[3] Brown, S. and Knight, P. *Assessing Learners in Higher Education*, Kogan Page (1994)
[4] Hopkins, K.D., Stanley, J.C. and Hopkins, B.R. (several editions) *Educational and Psychological Measurement and Evaluation*, Prentice Hall
[5] Linn, R.L. and Gronlund, (various editions with some title variation) *Measurement and Assessment in Teaching,* Merrill, Prentice-Hall
[6] Mehrens, W.A. and Lehmann, I.J. (various editions, all useful) *Measurement and Evaluation in Education and Psychology*, Holt, Rinehart & Winston
[7] Rowntree, D. (either edition) *Assessing Students: How Shall we Know Them?*, Harper & Row
[8] Thyne, J.M. *Principles of Examining*, (University of London Press 1974). Older but still useful.
[9] Wiersma, W. and Jurs, S.G. (either edition) *Educational Measurement and Testing*, Allyn & Bacon
[10] If your library uses Library of Congress cataloguing, look for titles at LB1131 or LB3051.