Integrative evaluation: An emerging role for classroom studies of CAL

This paper appeared in Computers and Education (1996) vol.26 no.1-3, pp.17-32

Stephen W. Draper, Margaret I. Brown, Fiona P. Henderson, Erica McAteer*

Department of Psychology

University of Glasgow

Glasgow G12 8QQ U.K.

email: steve@psy.gla.ac.uk

WWW: http://www.psy.gla.ac.uk/~steve

Abstract

This paper reviews the work of a team over two and a half years whose remit has been to "evaluate" a diverse range of CAL (Computer Assisted Learning) in use in a university setting. It gives an overview of the team's current method, including some of the instruments most often used, and describes some of the painful lessons from early attempts. It then offers a critical discussion of what the essential features of the method are, and of what such studies are and are not good for. One of the main conclusions, with hindsight, is that its main benefit is as "integrative evaluation": to help teachers make better use of the CAL by adjusting how it is used, rather than by changing the software or informing purchasing decisions.

1. Introduction

The authors constitute the evaluation group within a large project on CAL (Computer Assisted Learning) across a university, which has run for about two and half years at the time of writing [1]. We were charged with

* This paper describes work by the evaluation group within the TILT project (part of TLTP: see the acknowledgements). Enquiries about this paper (and other evaluation work) should be sent to the first author at the address above. Enquiries about TILT generally should be sent to the project director: g.doughty@elec.gla.ac.uk or G.F.Doughty, Robert Clark Centre, 66 Oakfield Avenue, Glasgow G12 8LS, U.K.

evaluating the CAL courseware whose use was being developed by other members of the project. In response we have performed over 20 studies of teaching software in our university across a very wide range of subject disciplines, from Dentistry to Economic History, from Music to Engineering, from Zoology practicals to library skills. More detailed accounts of some of these studies and of some of our methods are available elsewhere [2,3,4,5,6]. In this paper we review our experience as a whole, describe our current method, and discuss how it might be justified and what in fact it has turned out to be useful for.

2. General features of our approach

A popular term for the activity described here, and the one our project used for it, is "evaluation" — a term which implies making a (value) judgement. A better statement of our aim however is "to discover how an educational intervention performs" by observing and measuring the teaching and learning process, or some small slice of it. Our function is to provide better information than is ordinarily available about what is going on and its effects; it is up to others, most likely the teachers concerned, to use that information. One use would be to make judgements e.g. about whether to continue to use some piece of software ("summative" evaluation). Other uses might be to improve the design of the software ("formative" evaluation), to change how it is used (e.g. support it by handouts and tutorials — "integrative" evaluation), or to demonstrate the learning quality and quantity achieved (a QA function).

Our basic aim, then, was to observe and measure what we could about the courseware and its effects. Practical constraints would place some limits on what we could do towards this aim. Over time we learned a lot about how and how not to run such studies. Now we are in a position to review what uses studies under these conditions turn out to have: not exactly the uses we originally envisaged.

Our starting point was influenced by standard summative studies. We felt we should be able to answer questions about the effect of the courseware, and that this meant taking pre- and post-intervention measures. Furthermore, we felt these measures should derive from the students (not from on-lookers' opinions). This led to studies of the actual classes the software is designed for, and the need for instruments that can be administered to all students before and after the courseware. Such studies have the great virtue of being in a strong position to achieve validity: having realistic test subjects in real conditions.

Various pressures contribute to maintaining an emphasis on these special classroom opportunities, despite some drawbacks.

1. We are most likely to be invited to do a study (or, if we take the initiative, to secure agreement) when new courseware is being introduced to classes. There are several reasons for this.

1.1 This corresponds with many people's idea of evaluation as a one-shot, end of project, summative activity. Thus it is usually the default idea of developers, authors, teachers, funding bodies, universities etc.. The desirability of this is discussed below, but meanwhile it has a large effect on expectations, and hence on opportunities.

1.2 Experimental technique requires pre- and post-tests using quantitative measures in controlled conditions. This means that a lot of effort from the students, the teaching staff, and the investigators is put into each such occasion; and so it is unlikely that they can afford to do this often. Once a year is all that can be easily afforded.

2. The most important criterion in testing is whether the courseware is effective i.e. do students learn from it. It is hard to test this without a complete implementation.

3. It is important to get test subjects who are motivated to learn, and don't know the material already. These are often only available once a year in the target classes.

4. An advantage of this that we came to appreciate is that then the whole situation is tested, not just the courseware which in reality is only one factor in determining learning outcomes.

5. One important limit is the need not to overload the students. As investigators, we were happy to keep on multiplying the tests, questionnaires and interviews in order to learn more, but we quickly learned that students have a strictly limited tolerance for this addition to their time and trouble on such occasions. Hence potential instruments must compete with each other at minimising their cost to the students.

6. Investigator time is also a limiting factor. There are always far fewer investigators than students, so at best only a small sample can be interviewed and individually observed. We must therefore concentrate on paper instruments.

Thus our method focusses around such big occasions, and in effect is organised to make the most of these opportunities: to observe and measure what can be observed under these conditions. We tend to study real students learning as part of their course. We rely on paper instruments (e.g. questionnaires, multiple choice tests) to survey the whole class, supplemented by personal observations and interviews with a sample.

3. Overview of our "method"

Each particular study is designed separately, depending upon the goals of the evaluation, upon the particular courseware to be studied, and upon the teaching situation in which it is to be used. Each design draws on our battery of methods and instruments, which are still actively evolving, and which are selected and adapted anew for each study. In this section we describe what is common in our approach across most studies, but the considerable degree of individual adaptation means that the reader should understand that the term "method" should be read as having inverted commas: it has not been a fixed procedure mechanically applied. On the other hand, there has been a substantial uniformity in what we did despite great variations in the material being taught and in performing both formative evaluations of unfinished software and studies of finished software in its final classroom use.

Such studies are a collaborative effort between evaluators (ourselves), teachers (i.e. the lecturer or professor responsible for running the teaching activity plus his or her assistants), students, and (if the software is produced in-house) developers (the designers and writers of the software). Any failure to secure willing cooperation from the students was marked by a drop in the the number and still more in the usefulness of the returns. The teacher's cooperation is still more central: not only for permission and class time, but for the design of tests of learning gains, and interpretation of the results to which, as the domain experts, they are crucial contributors. As a rule, the evaluators will have most to say about the basic data, both evaluators and teachers will contribute substantially to interpretations of the probable issues and causes underlying the observations, and the teachers or developers will have most to say about what action if any they may now plan as a consequence.

Our method can be seen as having two levels. The "outer method" concerns collaborating with the teachers and developers to produce and agree a design for the study, carry it out at an identified opportunity, and produce a report including input from the teacher about what the observations might mean. Within that is an "inner method" for the selection and design of specific instruments and observational methods.

3.1 Our "outer method": interaction with teachers and developers

Generally we would follow a pattern summarised as follows:

1. One or more meetings, perhaps prepared for by a previous elicitation of relevant information by mail, of evaluators with teachers or developers to:

- Establish the teachers' goals for the evaluation
- Elicit the learning aims and objectives to be studied
- Elicit the classroom provision, course integration, and other features of the teaching situation and support surrounding the courseware itself e.g. how the material is assessed, whether it is a scheduled teaching event or open access.
- Establish options for classroom observation, questionnaire administration, interviews, etc.
- Discuss the feasibility of creating a learning quiz or other measure of learning gains.

2. An evaluator may go through the software to get an impression of what is involved there.

3. The teacher creates (if feasible) a quiz with answers and marking scheme, and defines any other assessment methods.

4. Evaluator and teacher finalise a design for the study.

5. Classroom study occurs.

6. A preliminary report from the evaluators is presented to the teacher, and interpretations of the findings sought and discussed.

7. Final report produced.

3.2 "Inner method": some instruments

Every study is different, but a prototypical design of a large study might use all of the following instruments.

Computer Experience questionnaire

This asks about past and current computer training, usage, skills, attitudes to and confidence about computers and computer assisted learning. We use this less and less, as has turned out to be seldom an important factor.

Task Experience Questionnaire

Where particular skills are targeted by courseware, it can be useful for teachers to have some information about students' current experience in the domain. If possible this questionnaire should be administered to students at a convenient point which is prior to the class within which the courseware evaluation itself is to run. (This is not to be confused with diagnostic tests to establish individual learning needs, which might be provided so that students can enter a series of teaching units at an appropriate level).

Observations (by evaluator, possibly using video)

Whenever possible we have at least one evaluator at a site as an observer, gathering open-ended observations. Sometimes we have set up a video camera to observe one computer station, while the human observers might watch another.

Student confidence logs (Example in appendix A.)

These are checklists of specific learning objectives for a courseware package, provided by the teacher, on which students rate their confidence about their grasp of underlying principles or their ability to perform tasks. Typically they do this immediately before encountering a set of learning materials or engaging in a learning exercise, then again immediately afterward. If there are several exposures to the information in different settings — tutorials, practical labs, lectures, independent study opportunities, say — then they may be asked to complete them at several times. A rise in scores for an item upon exposure to a particular activity shows that the students at least believe they have learned from it, while no rise makes it unlikely to have been of benefit. Since this instrument measures confidence rather than a behavioural demonstration of knowledge like an exam, it is only an indirect indication to be interpreted with some caution. Nevertheless, simple to use, these logs are proving to be an unexpectedly useful diagnostic instrument.

Knowledge quizzes (Example in appendix B.)

These are constructed by the teacher and given to students before and after an intervention, and at a later point (delayed post-test) where sensible. Each question usually corresponds to a distinct learning objective. For consistent marking, these quizzes are usually multiple choice. Their purpose is assessment of the courseware and other teaching — low average post-test scores on a particular question point to a learning objective that needs better, or different, treatment. High average pre-test scores may suggest that certain content is redundant. They can sometimes be usefully incorporated as a permanent part of the program, focusing attention before study, or for self assessment after a session.

Post Task Questionnaire (PTQ) (Example in appendix C.)

This is usually given immediately after the student has completed the class session which involves use of the courseware. It can be extended to fit other, related, classwork with optional questions depending on its occasion of use. It gathers "survey" information, at a fairly shallow level, about what students felt they were doing during the learning task, and why. It can also ask specific questions about points in which teachers, developers or evaluators are specifically interested — e.g. the use and usefulness of a glossary, perceived relevance of content or activity to the course in which it is embedded, etc. Also some evaluative judgements by students can be sought if wanted at that point. Ideally the information gained (which we are still seeking to extend as we develop the instrument further) should be supplemented by sample interviews and/or focus groups. The Resource Questionnaire (see below) shares some content items with the PTQ - where appropriate, questions could be loaded on to that, to prevent too much use of student time during class.

Focus groups or interviews with a sample of students

Focus groups consist of a group of students and an investigator, who will have a few set questions to initiate discussion. The students' comments act as prompts for each other, often more important than the investigator's starting questions. Both focus groups and interviews have two purposes. Firstly, as a spoken method of administering fixed questions (e.g. from the PTQ) they allow a check on the quality of the written responses from the rest of the students. Secondly they are used as an open-ended instrument to elicit points that were not foreseen when designing the questionnaires. Generally we get far more detail from these oral accounts than from written ones, especially if the answers are probed by follow up questions, and in addition we can ask for clarification on the spot of any unclear statements.

Learning Resource Questionnaire

With the teachers, we produce a checklist of available learning resources (e.g. lectures, courseware, books, tutorials, other students etc.) and ask students to indicate, against each item, whether they used it, how useful it was, how easily accessed, how critical for their study etc. We would normally administer this during a course, during exam revision, and perhaps after an exam. It has two main functions. The first is to look at students' independent learning strategies in the context of a department's teaching resources (which do they seek out and spend time on). The second is to evaluate those resources (which do they actually find useful). This questionnaire is especially useful for courses within which the computer package is available in open access labs, when it is not possible to monitor classes at a sitting by Post Task Questionnaires. This instrument is discussed in detail in [4].

Post Course questionnaire

Sometimes there are useful questions which can be asked of students at the end of a course or section of course which has involved the use of CAL. Teachers may seek information to expand or clarify unexpected observations made while the classes were in progress. We might want to get an overview from the students, to put with their daily "post task" feedback, or we may want to ask questions that may have seemed intrusive at the start of the course but, once students are familiar with course content, the evaluation itself and the evaluators, are more naturally answered. This instrument can be useful where a resource questionnaire (a more usual place for "post course" general questions) is not appropriate.

Access to subsequent exam performance on one or more relevant questions

In some cases it is later possible to obtain access to relevant exam results. Exam questions are not set to please evaluators, so only occasionally is there a clear correspondence between an exam item and a piece of CAL.

4. Some problematic issues

In the course of our studies, a number of issues struck us with some force. Often they were epitomised by a specific experience.

General points

Subjects with the right prior knowledge

One of us was called on to comment on a piece of CAL being developed on the biochemistry of lipid synthesis. As he had no background knowledge of biochemistry at all, he was inclined to suggest that the courseware was too technical and offered no handle to the backward students who might need it most. In fact, tests on the target set of students showed that this was quite wrong, and that they all were quite adequately familiar with the basic notation required by the courseware. This shows the importance, however informal the testing, of finding subjects with sufficient relevant pre-requisite knowledge for the material. It is not possible to hire subjects from the general population (not even the general undergraduate population) to test much of the CAL in higher education.

The converse point, perhaps more familiar but equally important, is that using subjects with too much knowledge is equally likely to be misleading. Hence asking teachers to comment is of limited value as they cannot easily imagine what it is like not to know the material already. Similarly, in one of our formative studies of a statistical package, we used students from a later point in the course who had already covered the material. On the few points where they had difficulty we could confidently conclude there was a problem to be cleared up, but when they commented that on the whole they felt the material was too elementary we could not tell whether to accept that as a conclusion or that it was simply because they already knew the material. This illustrates how, while some conclusions can be drawn from other subjects (if over-qualified subjects have difficulty or under-qualified subjects find it trivial then there must be a problem), only subjects with exactly the right prior knowledge qualifications are most useful. This contributes to the pressure on using real target classes, despite the limited supply.

The evaluators' dependence on collaboration with teachers

The same experiences and considerations also show that evaluators are often wholly unqualified in domain knowledge especially for university courses, whether of biochemistry or of sixteenth century musicianship. Teachers are qualified in the domain knowledge, and so evaluators must depend on them for this (e.g. in designing knowledge quizzes, and in judging the answers), as well as for interpreting findings (whether of poor scores or of student comments) in terms of what the most likely causes and remedies are. On the other hand teachers are over-qualified in knowledge, hence the value of doing studies with real students.

Subjects with the right motivation to learn

Subjects must equally have the right motivation to learn (i.e. the same as in the target situation). Where would you find subjects naturally interested in learning about statistics or lipid synthesis? Such people are rare, and in any case not representative of the target group, who in many cases only learn because it is a requirement of the course they are on. This was brought home to us in an early attempt to study a Zoology simulation package. The teacher announced its availability and recommended it to a class. The evaluators turned up, but not a single student. Only when it was made compulsory was it used.

Paying subjects is not in general a solution, although it may be worth it for early trials, where some useful results can be obtained without exactly reproducing the target conditions. Sometimes people are more motivated by helping in an experiment than they are by other factors. For instance if you say you are doing research on aerobic exercise you may easily persuade people to run round the block, but if you ask your friends to run round the block because watching them sweat gives you pleasure you are more likely to get a punch on the nose than compliance. In other words, participation in research may produce more motivation than friendship can. On the other hand with educational materials, paid subjects may well process the material politely but without the same kind of attempt to engage with it that wanting to earn a qualification may produce. In general, as the literature on cognitive dissonance showed, money probably produces a different kind of motivation. The need for the right motivation in subjects, like the need for the right level of knowledge, is a pressure towards using the actual target classes in evaluation studies.

Measures

Attitude measures

Asking students at the end of an educational intervention (EI — for example a piece of courseware) what they feel about it is of some interest, as teachers would prefer that students enjoy courses. However learning gains are a far more important outcome in most teachers' view, and attitudes are very weak as a measure of that educational effect. Attitudes seem to be mainly an expression of how the experience compared to expectations. With CAL, at least currently, attitudes vary widely even within a class but above all across classes in different parts of the university. For instance we have observed some student groups who had the feeling that CAL was state of the art material that they were privileged to experience, and other groups who viewed it as a device by negligent staff to avoid the teaching which the students had paid for. In one case, even late in a course of computer mediated seminars that in the judgement of the teaching staff and on measures of content and contributions was much more successful than the non-computer form they replaced, about 10% of the students were still expressing the view that if they had known in advance that they would be forced to do this they would have chosen another course. Thus students express quite strongly positive and negative views about a piece of courseware that often seem unrelated to their actual educational value. This view of attitudes being determined by expectation not by real value is supported by the corollary observation that when the same measures are applied to lectures (as we did in a study that directly compared a lecture with a piece of CAL), students did not express strong views. In followup focus groups, however, it emerged that they had low expectations of lectures, and their experience of them was as expected; whereas the CAL often elicited vocal criticisms even though they learned successfully from it, and overall brought out many more comments both positive and negative reflecting a wide range of attitudes, presumably because prior expectations were not uniform in this group.

Measuring the shift in attitude instead (i.e. replacing a single post-EI attitude measure by pre- and post-measures of attitude) does not solve the problem. A big positive shift could just mean that the student had been apprehensive and was relieved by the actual experience, a big negative shift might mean they had had unrealistic expectations, and no shift would mean that they had had accurate expectations but might mean either great or small effectiveness.

These criticisms of attitude measures also apply in principle to the course feedback questionnaires now becoming widespread in UK universities. As has been shown (e.g. [7]), these have a useful level of validity because, we believe, students' expectations are reasonably well educated by their experience of other courses: so a course that is rated badly compared to others probably really does have problems. However students' widely varying expectations of CAL as opposed to more familiar teaching methods render these measures much less useful in this context, increasing the importance of other measures. Attitudes are still important to measure, however, as teachers are likely to want to respond to them, and perhaps attempt to manage them.

Confidence logs

Confidence logs ask students, not whether they thought an EI was enjoyable or valuable, but whether they (now) feel confident of having attained some learning objective. Like attitude measures, they are an indirect measure of learning whose validity is open to question. They have been of great practical importance however because they take up much less student time than the quizzes that measure learning more directly, and so can be applied more frequently e.g once an hour during an afternoon of mixed activities.

By and large confidence logs can be taken as necessary but not sufficient evidence: if students show no increase in confidence after an EI it is unlikely that they learned that item on that occasion, while if they do show an increase then corroboration from quiz scores is advisable as a check against over confidence on the students' part. Even the occasional drop in confidence that we have measured is consistent with this view: in that case, a practical exercise seems to have convinced most students that they had more to learn to master the topic than they had realised. We have often used confidence logs several times during a long EI, and quizzes once at the end. In these conditions, the quizzes give direct evidence about whether each learning objective was achieved, while the confidence logs indicate which parts of the EI were important for this gain.

Knowledge quizzes

Although knowledge quizzes are a relatively direct measure of learning, they too are subject to questions about their validity. When scores on one item are low this could be due either to poor delivery of material for that learning objective or because the quiz item was a poor test of it or otherwise defective. When we feed back the raw results to the teacher they may immediately reconsider the quiz item (which they themselves designed), or reconsider the teaching. Although this may sound like a methodological swamp when described in the abstract, in practice such questions are usually resolved fairly easily by interviewing the students about the item. Thus our studies are not much different from other practical diagnostic tasks such as tracking down a domestic electrical fault: it is always possible that the "new" light bulb you put in was itself faulty, but provided you keep such possibilities in mind it is not hard to run cross checks that soon provide a consistent story of where the problem or problems lie.

Much of the usefulness of both quizzes and confidence logs comes from their diagnostic power, which in turn comes from the way they are specific to learning objectives, each of which is tested as a separate item. For instance when one item shows little or no gain compared to the others, teachers can focus on improving delivery for that objective. Such differential results simultaneously give some confidence that the measures are working (and not suffering from ceiling or floor effects), that much of the learning is proceeding satisfactorily, that there are specific problems, and where those problems are located. In other words they both give confidence in the measures and strongly suggest (to the teachers at least) what might be done about the problems. In this respect they are much more useful than measures applied to interventions as a whole (such as course feedback questionnaires), which even when their results are accepted by teachers as indicating a problem (which they often are not), are not helpful in suggesting what to change.

Delayed learning gains

In many educational studies, measures of learning outcomes are taken not only before and immediately after the EI, but also some weeks or months later at a delayed post-test. This is usually done to measure whether material, learned at the time, is soon forgotten. In higher education, however, there is an opposite reason: both staff and students often express the view that the important part of learning is not during EIs such as lectures or labs, but at revision time or other self-organised study times. If this is true, then there might be no gains shown at an immediate post-test, only at a delayed test.

Auto compensation

This all seems to make it clear that the best measure would be the difference between pre-test and delayed post-test. However there is an inescapable problem with this: that any changes might be due to other factors, such as revision classes, going to text books etc. Indeed this is very likely, both from external factors (other activities relating to the same topic), and also from a possible internal factor that might be called "auto compensation". In higher education students are given a lot of responsibility for their learning, and how they go about it. This means that if they find one resource unsatisfactory or unlikeable, they are likely to compensate by seeking out a remedial source. Thus, for instance, bad lectures may not cause bad exam results because the students will go to textbooks, and similarly great courseware may not cause better exam results but simply allow students to get the same effect with less work from books.

Thus although we gather delayed test results (e.g. from exams) when we can, these really only give information on the effect of the course as a whole including all resources. Only immediate post-tests can pick out the effects of a specific EI such as a piece of CAL, even if important aspects of learning and reflection may only take place later.

Open-ended measures

All the above points have been about measures that are taken systematically across the whole set of subjects and can be compared quantitatively. More important than all of them however are the less formal open-ended measures (e.g. personal observations, focus groups, interviews with open-ended questions). This is true firstly because most of what we have learned about our methods and how they should be improved has come from such unplanned observations: about what was actually going on as opposed to what we had planned to measure. Not only have we drawn on them to make the arguments here about big issues, but also they have often been important in interpreting small things. For instance in one formative study there was really only one serious problem where all the students got stuck, but not all of them commented on this clearly in the paper instruments. It was only our own observation that made this stand out clearly, and so made us focus on the paper measures around this point. Open-ended measures are also often valued by the teachers: transcribed comments are a rich source of insight to them.

Other factors affecting learning

Our studies were often thought of as studies of the software and its effects. The problem with this view was brought home to us in an early study of Zoology simulation software being used in a lab class as one "experiment" out of about six that students rotated between. We observed that the teacher running the class would wait until students at the software station had finished their basic run through and then engage them at this well chosen moment in a tutorial dialogue to bring out the important conceptual points. Obviously any learning gains recorded could not be simply ascribed to the software: they might very well depend on this human tutoring. This was not anything the documentation had suggested, nor anything the teacher had mentioned to us: it was basically unplanned good teaching. On the other hand, the teacher's first priority in being there was to manage the lab i.e. handling problems of people, equipment, materials would have had first priority, so there was no guarantee of these tutorial interactions being delivered.

This showed that our studies could not be thought of as controlled experiments, but had the advantages and disadvantages of realistic classroom studies. We could have suppressed additional teacher input and studied the effect of the software by itself, but this would have given the students less support and in fact been unrealistic. We could have insisted on the tutoring as part of the EI being studied. This is sometimes justified as "using the software strictly in the way for which it was designed". But there are several problems with this: firstly, having a supervising teacher free for this is not something that could be guaranteed, so such a study would, like excluding tutoring, achieve uniform conditions at the expense of studying real conditions (it would have required twice the staff: one to manage the lab, one to do the tutorials). Secondly, in fact we would not have known that this was the appropriate condition to study, as neither the designers nor the teacher had planned this in advance: like a lot of teaching, such good practice is not a codified skill available for evaluators to consult when designing experiments, but rather something that may be observed and reported by them.

Our studies, then, observe and can report real classroom practice. They are thus "ecologically valid". They cannot be regarded as well controlled experiments. On the other hand, although any such experiment might be replicable in another experiment, it would probably not predict what results would be obtained in real teaching. Our studies cannot be seen as observing "the effect of the software", but rather the combined effect of the overall teaching and learning situation. Complementary tutoring was one additional factor in this situation. Another mentioned earlier was external coercion to make students use the software at all (a bigger factor in determining learning outcomes than any feature of the software design). Some others should also be noted.

Study habits for CAL

In one study, a class was set to using some courseware. After watching them work for a bit, the teacher saw that very few notes were being taken and said "You could consider taking notes." Immediately all students began writing extensively. After a few minutes she couldn't bear it and said "You don't have to copy down every word", and again every students' behaviour changed. These students were in fact mature students, doing an education course, and in most matters less suggestible than typical students and more reflective about learning methods. Their extreme suggestibility here seems to reflect that, in contrast to lectures, say, there is no current student culture for learning from CAL from which they can draw study methods. The study habits they do employ are likely to have large effects on what they learn, but cannot be regarded as either a constant of the situation (what students will normally do) or as a personality characteristic. From the evaluation point of view it is an unstable factor. From the teaching viewpoint, it is probably something that needs active design and management. Again, it is a factor that neither designer nor teacher had planned in advance.

Hawthorne effects

The Hawthorne effect is named after a study of work design in which whatever changes were made to the method of working, productivity went up, and the researcher concluded that the effect on the workers of being studied, and the concern and expectations that that implied, were having a bigger effect than changes to the method in themselves. Applied to studies of CAL, one might in principle distinguish between a Hawthorne effect of the CAL itself, and a Hawthorne effect of the evaluation. As to the latter, we have not seen any evidence of students being flattered or learning better just because we were observing them. In principle one could test this by comparing an obtrusive evaluation with one based only on exams and other class assessment, but we have not attempted this. Perhaps more interesting is the possibility, also not yet tested, that evaluation instruments such as confidence logs may enhance learning by promoting reflection and self-monitoring by students. Should this be the case, such instruments could easily become a permanent part of course materials (like self-assessment questions) to maintain the benefit.

As to the effect from the CAL i.e. of learning being improved because students and perhaps teachers regard the CAL as glamourous advanced material, this is probably often the case, but we expect there equally to be cases of negative Hawthorne effects, as some groups of students regard it from the outset as a bad thing. We cannot control or measure this effect precisely: unlike studies of medicines but like studies of surgical procedures, it is not possible to prevent subjects from knowing what "treatment" they are getting, and any psychological effects of expectations to be activated. However it does mean that measuring students' attitudes to CAL should probably be done in every study, so that the likely direction of any effect is known. (This implies, in fact, that the most impressive CAL is that which achieves good learning gains even with students who don't like it — clearly that CAL is having more than a placebo effect.)

5. Summary of the core attributes of our approach

Our work began by trying to apply a wide range of instruments. We then went through a period of rapid adjustment to the needs and constraints of the situations in which our opportunities to evaluate occurred. In earlier sections we have described the components of our method, and then the issues which impressed themselves upon us as we adapted. What, in retrospect, are the essential features of our resulting method?

Our approach is empirical: based on observing learning, not on judgements made by non-learners, expert or otherwise. Our studies have usually been felt to be useful by the teachers, even when we ourselves have felt that the issues were obvious and did not warrant the effort of a study: simply presenting the observations, measures, and collected student comments without any comment of our own has had a much greater power to convince teachers and start them actively changing things than any expert review would have had.

This power of real observations applies equally to ourselves: we have in particular learned a great amount from open-ended measures and observations, which is how the unexpected can be detected. We therefore always include them, as well as planning to measure pre-identified issues. The chief of the latter are learning gains, which we always attempt to measure. Our measures for this (confidence logs and quizzes) are related directly to learning objectives, which gives these measures direct diagnostic power.

Tests on CAL in any situation other than the intended teaching one are only weakly informative, hence our focus on classroom studies. This is because learning, especially in higher education, depends on the whole teaching and learning situation, comprising many factors, not just the "material" (e.g. book or courseware). You have only to consider the effects on learning of leaving a book in a pub for "free access use", versus using it as a set text on a course, backed up by instructions on what to read, how it will be tested, and tutorials to answer questions. In any study there is a tension between the aims of science and education, between isolating factors and measuring their effects separately, and observing learning as it normally occurs with a view to maximising it in real situations. Our approach is closer to the latter emphasis. Partly because of this, these studies are crucially dependent on collaboration with the teacher, both for their tacit expertise in teaching and their explicit expertise in the subject being taught, which are both important in interpreting the observations and results.

Ours is therefore distinct from various other approaches to evaluation. It differs from the checklist approach (e.g. [8]) because it is "student centered" in the sense that it is based on observation of what students actually do and feel. It differs from simply asking students their opinion of a piece of CAL because it attempts to measure learning, and to do this separately for each learning objective. It differs from simple pre- and post-testing designs because the substantial and systematic use of open-ended observation in real classroom situations often allows us to identify and report important factors that had not been foreseen.

It also differs from production oriented evaluation, geared around the production of course material. These are similar in spirit to software engineering approaches in that they tend to assume that all the decisions about how teaching is delivered are "design decisions" and either are or should be taken in advance by a remote design team and held fixed once they have been tested and approved. In contrast our approach is to be prepared to observe and assist in the many local adaptations of how CAL is used that occur in practice, and to evaluate the local situation and practice as much as the CAL. It is our experience that designers of CAL frequently do not design in detail the teaching within which its use will be embedded, that teachers make many local adaptations just as they do for all other teaching, and that even if designers did prescribe use many teachers would still modify that use, just as they do that of textbooks.

The food industry provides an analogy to this contrast. Prepared food ready for reheating is carefully designed to produce consistent results to which the end user contributes almost nothing. The instructions on the packaging attempt to ensure this by prescribing their actions. Production oriented evaluation is appropriate for this. In contrast, supplying ingredients (meat, vegetables, etc.) for home cooking involves rather different issues. Clearly studies of how cooks use ingredients are important to improving supplies. However while some meals such as grilled fish and salads depend strongly on the quality of the ingredients, others such as pies, stews and curries were originally designed to make the best of low quality ingredients, and remain so successful that they are still widely valued. Such recipes are most easily varied for special dietry needs, most dependent on the cook's skills, and most easily adapted depending on what ingredients are available. Production oriented evaluation, organised as if there were one ideal form of the recipe, is not appropriate: rather an approach like ours that studies the whole situation is called for.

6. Discussion: what is the use of such studies in practice?

Having described and discussed many aspects of our approach, and having tried to summarise its core features, we now turn to the question of what, in retrospect, the uses of studies like ours turn out to be. We consider in turn five possible uses: formative, summative, illuminative, integrative, and QA functions.

6.1 Formative evaluation

"Formative evaluation" is testing done with a view to modifying the software to solve any problems detected. Some of our studies have been formative, and contributed to improvements before release of the software. Because the aim is modification, the testing needs not only to detect the existence of a problem (symptom detection) but if possible to suggest what modification should be done: it needs to be diagnostic and suggestive of a remedy. Open-ended measures are therefore vital here to gather information about the nature of any problem (e.g. do students misunderstand some item on the screen, or does it assume too much prior knowledge, or what?). However our learning measures (quizzes and confidence logs) are also useful here because, as they are mostly broken down into learning objectives, they are diagnostic and indicate with which objective the problem is occurring. In formative applications we might sharpen this further by asking students after each short section of the courseware (which would usually correspond to a distinct learning objective) to give ratings for the presentation, for the content, and for their confidence about a learning objective (the one the designers thought corresponded to that section), and also perhaps answer quiz items.

However many problems cannot be accurately detected without using realistic subjects and conditions, for reasons given earlier. Both apparent problems and apparent successes may change or disappear when real students with real motivations use the courseware. By that time however the software is by definition in use, and modifications to future versions of the software may not be the most important issue.

6.2 Summative evaluation

"Summative evaluation" refers to studies done after implementation is complete, to sum up observed performance in some way, for instance consumer reports on microwave ovens. It could be used to inform decisions about which product to select, for example. They can be multi-dimensional in what they report, but in fact depend on there being a fairly narrow definition of how the device is to be used, and for what. As we have seen, this does not apply to the use of CAL in higher education. What matters most are the overall learning outcomes, but these are not determined only or mainly by the CAL: motivation, coercion, and other teaching and learning activities and resources are all crucial and these vary widely across situations. It is probably not sensible in practice to take the view that how a piece of courseware is used should be standardised, any more than a textbook should only be sold to teachers who promise to use it in a particular way.

It is not sensible to design experiments to show whether CAL is better than lectures, any more than whether textbooks are good for learning: it all depends on the particular book, lecture, or piece of CAL. Slightly less obviously, it is not sensible to run experiments to test how effective a piece of CAL is, because learning is jointly determined by many factors and these vary widely across the situations that occur in practice. Well controlled experiments can be designed and run, but their results cannot predict effectiveness in any other situation than the artificial one tested because other factors have big effects. This means that we probably cannot even produce summative evaluations as systematic as consumer reports on ovens. Ovens are effectively the sole cause of cooking for food placed in them, while CAL is probably more like the role of an open window in cooling a room in summer: crucial, but with effects that depend as much on the situation as they do on the design of the window.

However this does not mean that no useful summative evaluation is possible. When you are selecting a textbook or piece of CAL for use you may have to do this on the basis of a few reviews, but you would certainly like to hear that someone had used it in earnest on a class, and how that turned out. The more the situation is like your own, the more detailed measures are reported, and the more issues (e.g. need for tutorial support) identified as critical, the more useful such a report would be. In this weak but important sense, summative evaluations of CAL are useful. Many of our studies have performed summatively in this way, allowing developers to show enquirers that the software has been used and tested, and with substantial details of its performance.

6.3 Illuminative evaluation

"Illuminative evaluation" is a term introduced by Parlett & Hamilton [9, 10] to denote an observational approach inspired by ethnographic rather than experimental traditions and methods. Its aim is to discover, not how an EI (educational intervention) performs on standard measures, but what the factors and issues are that are important to the participants in that particular situation, or which seem evidently crucial to a close observer.

The importance of this has certainly impressed itelf upon us, leading to our stress on the open-ended methods that have taught us so much. In particular, they allow us to identify and report on factors important in particular cases, which is an important aspect of our summative reports, given the situation-dependent nature of learning. They have also allowed us to identify factors that are probably of wider importance, such as the instability but importance of the study methods that students bring to bear on CAL material.

Thus our studies have an important illuminative aspect and function, although they combine it with systematic comparative methods, as Parlett & Hamilton originally recommended. Whether we have achieved the right balance or compromise is harder to judge.

6.4 Integrative evaluation

Although our studies can and have performed the traditional formative and summative functions discussed above, in a number of cases we have seen a new role for them emerge of perhaps greater benefit. The experience for a typical client of ours (i.e. someone responsible for a course) is of initially low expectations of usefulness for the evaluation — after all, if they had serious doubts about whether the planned teaching was satisfactory they would already have started to modify it. However when they see the results of our evaluation, they may find on the one hand confirmation that many of their objectives are indeed being satisfactorily achieved (and now they have better evidence of this than before), and on the other hand that some did unexpectedly poorly — but that they can immediately think of ways to tackle this by adjusting aspects of the delivered teaching without large costs in effort or other resources. For example, a particular item shown to be unsuccessfully handled by the software alone might be specifically addressed in a lecture, supplemented by a new handout, or become the focus of companion tutorials. This is not very different from the way that teachers dynamically adjust their teaching in the light of other feedback e.g. audience reaction, and is a strength of any face to face course where at least some elements (e.g. what is said, handouts, new OHP slides) can be modified quickly. The difference is in the quality of the feedback information. Because our approach to evaluation is based around each teacher's own statement of learning objectives, and around the teacher's own test items, the results are directly relevant to how the teacher thinks about the course and what they are trying to achieve: so it is not surprising that teachers find it useful.

An example of this is a recent study of automated psychology labs, where students work their way through a series of computer-mediated experiments and complete reports in an accompanying booklet. In fact the objectives addressed by the software were all performing well, but the evaluation showed that the weakest point was in writing the discussion section of the reports, which was not properly supported. This focussed diagnosis immediately suggested (to the teacher) a remedy that will be now be acted on (a new specialised worksheet and a new topic for tutorials), where previous generalised complaints had not been taken as indicating a fault with the teaching.

How does this kind of evaluation fit in with the other kinds of available feedback about teaching? The oldest and probably most trusted kind of feedback in university teaching is direct verbal questions, complaints, and comments from students to teachers. Undoubtedly many local problems are quickly and dynamically adjusted in this way. Its disadvantages, which grow greater with larger class sizes, are that the teacher cannot tell how representative of the class each comment is, and that obviously "typical" students do not comment because only a very small, self-selected, number of students get to do this. Course feedback questionnaires get round this problem of representativeness by getting feedback from all students. However they are generally used only once a term, and so are usually designed to ask about general aspects of the course, such as whether the lecturer is enthusiastic, well-organised, and so on. It is not easy for teachers to see how to change their behaviour to affect such broad issues, which are certainly not directly about specific content (which after all is the whole point of the course), how well it is being learned, and how teachers could support that learning better. Our methods are more detailed, but crucially they are much more diagnostic: much more likely to make it easy for teachers to think of a relevant change to address the problem.

This constitutes a new role for evaluation that may be called "integrative": evaluation aimed at improving teaching and learning by better integration of the CAL material into the overall situation. It is not primarily either formative or summative of the software, as what is both measured and modified is most often not the software but surrounding materials and activities. It is not merely reporting on measurements as summative evaluation is, because it typically leads to immediate action in the form of changes. It could therefore be called formative evaluation of the overall teaching situation, but we call it "integrative" to suggest the nature of the changes it leads to. This role for evaluation is compatible with the issues that are problems for the role of summative evaluation, such as observing only whole situations and the net effect of many influences on learning. After all, that is what teachers are in fact really concerned with: not the properties of CAL, but the delivery of effective teaching and learning using CAL.

6.5 QA functions

Such integrative evaluation can also be useful in connection with the QA (quality audit, assessment, or assurance) procedures being introduced in UK universities, and in fact can equally be applied to non-CAL teaching. Firstly it provides much better than normal evidence about quality already achieved. Secondly it demonstrates that quality is being actively monitored using extensive student-based measures. Thirdly, since it usually leads to modifications by the teachers without any outside prompting, it provides evidence of teachers acting on results to improve quality. Thus performing integrative evaluations can leave the teachers in the position of far exceeding current QA standards, while improving their teaching in their own terms.

These advantages stem from our adoption of the same objectives-based approach that the positive side of QA is concerned with. In practice it can have the further advantage that teachers can use the evaluation episode to work up the written statement of their objectives in a context where this brings them some direct benefit, and where the statements get debugged in the attempt to associate them with test items. They can then re-use them for any QA paperwork they may be asked for at some other time. This can overcome the resistance many feel when objectives appear as a paper exercise divorced from any function contributing to the teaching.

6.6 Limitations and the need for future work

Our approach has various limitations associated with the emphasis on particular classroom episodes. We have not developed convincing tests of deep as opposed to shallow learning [11]: of understanding as opposed to the ability to answer short quiz items. Thus we have almost always looked at many small learning objectives, rather than how to test for large ones concerning the understanding of central but complex concepts. This should not be incompatible with our approach, but will require work to enable us to suggest to our clients how to develop such test items. Similarly we have considered, but not achieved, measures of a student's effective intention in a given learning situation (their "task grasp") which probably determines whether they do deep learning, shallow learning or no learning. For instance, when a student is working through a lab class is she essentially acting to get through the afternoon, to complete the worksheet, to get the "right" result, or to explore scientific concepts? The same issue is important in CAL, and probably determines whether students flip through the screens, read them through once expecting learning to just happen, or actively engage in some definite agenda of their own.

It is also important to keep in mind a more profound limitation of the scope of such studies: they are about how particular teaching materials perform, and how to adjust the overall situation to improve learning. Such studies are unlikely to initiate big changes and perhaps big improvements such as a shift from topic-based to problem-based learning, or the abandonment of lectures in favour of other learning activities. They will not replace the research that goes into important educational advances, although they can, we believe, be useful in making the best of such advances by adjusting their application when they are introduced. Integrative evaluation is likely to promote small and local evolutionary adaptations, not revolutionary advances.

Acknowledgements

This work was done as part of the TILT (Teaching with Independent Learning Technologies) project, funded through the TLTP programme (Teaching and Learning Technology Programme) by the UK university funding bodies (DENI, HEFCE, HEFCW, SHEFC) and by the University of Glasgow. The studies discussed here could

not have been done without the active participation of many members of the university teaching staff to whom we are grateful.

References

1. Doughty,G., Arnold,S., Barr,N., Brown,M.I., Creanor,L., Donnelly,P.J., Draper,S.W., Duffy,C., Durndell,H., Harrison,M., Henderson,F.P., Jessop,A., McAteer,E., Milner,M., Neil,D.M., Pflicke,T., Pollock,M., Primrose,C., Richard,S., Sclater,N., Shaw,R., Tickner,S., Turner,I., van der Zwan,R. & Watt,H.D. Using learning technologies: interim conclusions from the TILT project TILT project report no.3, Robert Clark Centre, University of Glasgow (1995)

2. Draper,S.W., Brown,M.I., Edgerton,E., Henderson,F.P., McAteer,E., Smith,E.D., & Watt,H.D. Observing and measuring the performance of educational technology TILT project report no.1, Robert Clark Centre, University of Glasgow (1994)

3. Creanor,L., Durndell,H., Henderson,F.P., Primrose,C., Brown,M.I., Draper,S.W., McAteer,E. A hypertext approach to information skills: development and evaluation TILT project report no.4, Robert Clark Centre, University of Glasgow (1995)

4. Brown,M.I., Doughty,G.F., Draper,S.W., Henderson,F.P., McAteer,E. "Measuring learning resource use" submitted to this issue of Computers & Education (1996)

5. McAteer,E., Brown,M.I., Draper,S.W., Henderson,F.P., Barr,N. & Neil,D. "Simulation software in a life sciences practical laboratory" submitted to this issue of Computers & Education (1996)

6. Henderson,F.P., Creanor,L., Duffy,C. & Tickner,S. "Case studies in evaluation" submitted to this issue of Computers & Education (1996)

7. Marsh, H.W. "Student's evaluations of university teaching: research findings, methodological issues, and directions for future research" Int. journal of educational research vol.11 no.3 pp.253-388 (1987)

8. Machell, J. & Saunders,M. (eds.) MEDA: An evaluation tool for training software Centre for the study of education and training, University of Lancaster (1991)

9. Parlett, M.R. & Hamilton,D. "Evaluation as illumination: a new approach to the study of innovatory programmes".
Workshop at Cambridge, and unpublished report Occasional paper 9, Centre for research in the educational sciences, University of Edinburgh. (1972)
And in D.Hamilton, D.Jenkins, C.King, B.MacDonald & M.Parlett (eds.) Beyond the numbers game: a reader in educational evaluation (Basingstoke: Macmillan) ch.1.1 pp.6-22. (1977)
And in R.Murphy & H.Torrance (eds.) Evaluating education: issues and methods (Milton Keynes: Open University Press) ch.1.4 pp.57-73 (1987)

10. Parlett, M. & Dearden,G. Introduction to illuminative evaluation: studies in higher education (Pacific soundings press) (1977)

11. Marton,F., Hounsell,D. & Entwistle,N. (eds.) The experience of learning (Edinburgh: Scottish academic press) (1984)

Appendix A: Student confidence logs

Please indicate, by ticking the relevant box, how confident you feel that you are able to:

Topic	Very Confident	Confident	Some Confidence	Little Confidence	No Confidence Whatsoever
a) Sequence text with sound					
b) Accurately identify the keyword in a given statement or question					
c) Describe the probabilisitic behaviour of channel gates					
d) Provide antibiotic prophylaxis for patients allergic to penicillin					

These statements are from four different exercises evaluating:

(a) an authoring software tutorial for HE teachers learning to produce simple multimedia courseware for their students.

(b) from a study skills package for first year undergraduates, on library search strategies.

(c) simulation courseware provided to 3rd year Life Science students during a Practical Lab course in Animal Physiology and Neurobiology.

(d) from an evaluation of different teaching components (tutorials, demonstrations, practical work and courseware) within a 3rd year practical course in periodontology.

Appendix B: Knowledge quizzes

The following example questions, provided by Maggie Pollock, were used during the evaluation of TILT-funded courseware on the principles of Impulse and Momentum.

1. Momentum is defined as:-

   - mv^2
   - m^2v
   - mv
   - 2mv
   - I don't know

3. The Impulse of a Force is defined to be the:-

   - Change in Velocity
   - Change in Momentum
   - Change in Kinetic Energy
   - Change in Mass
   - I don't know

8. If a collision is perfectly elastic, which of the following statements are true:-

   - The bodies separate after impact
   - The bodies stick together after impact
   - The value of e=0
   - The value of e=1
   - None of these
   - I don't know

Appendix C: Post Task Questionnaire

Excerpt from PTQ used for a language learning package

Course ____________________ **Matric. Number** [| | | | | | |]

Read each question carefully. Answer all relevant questions as accurately as you can.
Please PRINT your written answers.

2) Please list the things you particularly liked and you particularly disliked about the package:-

Liked:

Disliked:

3) What did you think you were meant to learn from this package?
Please list below:

4) Did you? Yes No Some of them
Please comment further if you wish to:

5) Did you feel that the work you did in this Computer package episode complemented what you covered in this week's language class? Yes No
Please comment further if you wish to:

6) What did you spend most of your time trying to do during this computer session?
Please indicate by circling a point on the following scale:

Discovering details of how to operate the program	• • • 50/50 • • •	Concentrating on subject-related problems and answers

Comments: (please feel free to make any comments, positive or negative, that you would like to add to this form, about the computer package, or the course as a whole, or the way that we are evaluating it - i.e. the questionnaires!)