Prospects for summative evaluation

of CAL in higher education

Stephen W. Draper

Department of Psychology, University of Glasgow

Many developers and evaluators feel an external demand on them for summative evaluation of courseware. Problems soon emerge. One is that the CAL may not be used at all by students if it is not made compulsory. If one measures learning gains, how does one know that one is measuring the effect of the CAL or of the motivation in that situation? Such issues are the symptoms of the basic theoretical problem with summative evaluation, which is that CAL does not cause learning like turning on a tap, any more than a book does. Instead, it is one rather small factor in a complex situation. It is of course possible to do highly controlled experiments: for example to motivate the subjects in a standardized way. This should lead to measurements that are repeatable by other similar experiments. However they will be measurements that have little power to predict the outcome when the CAL is used in real courses. Hence the simple view of summative evaluation must be abandoned. Yet it is possible to gather useful information by studying how a piece of CAL is used in a real course and what the outcomes were. Although this does not guarantee the same outcomes for another purchaser, it is obviously useful to know that the CAL has been used successfully one or more times, and how it was used on those occasions. Such studies can also serve a different 'integrative' rather than summative function by pointing out failings of the CAL software and suggesting how to remedy them.

Introduction

Summative evaluation is evaluation done after software design and production is complete in order to establish its performance and properties. A prototypical case would be the tables produced in the consumer magazine *Which?* comparing a considerable range of properties of alternative available products (for example, washing machines) that they have measured in their own trials. Thus summative evaluation is not only done after production; it is typically about comparative measurements done to assist decisions concerning purchase.

Many developers and evaluators feel an external demand on them for summative evaluation of courseware. They feel they are being asked to prove that the software 'works', to show that it is cost-effective, that it is durable, and that it is worth the price to the purchaser. This is seen as a matter of testing the software, as in most software-development projects. Thus tests are done by using the software, and measuring various

outcomes of its use (for example, how people think about it, what is learned). Sometimes the software's performance is compared with some alternative such as no software or traditional teaching.

However, we know much less about the ingredients of successful teaching delivery than we do about washing clothes, and this has important consequences for what we can learn from measurements and what we in fact want to find out.

Symptoms of problems with the obvious approach

In our work on the TILT project (Doughty *et al*, 1995), we were soon struck by features that cast doubt on the sense of our doing evaluation of this kind (Draper *et al*, 1994).

1. The first is that the CAL may not be used at all by students if it is not made compulsory (here I shall use the term 'CAL' to refer indiscriminately to any computer software that might be introduced by teachers to support learning). This draws attention to the crucial role of motivation. If you measure learning gains, how do you know you are measuring the effect of the CAL or of the motivation in that situation? Certainly in the case mentioned, the CAL alone produced no learning because it produced no usage: motivation created by a teacher was crucial.

2. Another issue is that of the actions of teachers, for instance engaging students in a Socratic dialogue based around the CAL software. Obviously, any learning gains would be affected, and probably dominated, by the teacher's skill. But to evaluate software in the absence of teachers is to measure a different situation from the most common one in higher education, and furthermore one that would not get the most from the software; hence it is neither realistic (valid) nor constructive.

3. We have observed student study strategies such as note-taking radically changed by short remarks by the teacher, far more so than in a lecture. At least at the present time, it seems that CAL does not usually elicit a stable study strategy, while teachers can and do influence it in a big way. Since study strategies have a large effect on learning outcomes, again it seems beside the point to look for measurements independent of these; rather, the point would be to discover which study strategy is best for each piece of courseware and how to ensure that students adopt it.

4. Although some software is designed to be used once and never referred to again, much courseware is like textbooks and is intended to serve as reference and revision material as well as, or instead of, primary exposition. That means that the relevant tests of learning must be delayed until after the examination. However, it is then hard to tell how much, if at all, the students depended on the courseware as opposed to alternative resources such as books. Any evaluation will be indicative not about the properties of the software but about how the overall set of resources and student activities performed.

5. As a corollary to this, if in fact students find the courseware useless, they are likely to compensate by relying more on alternative resources (we might call such self-monitoring and correction 'auto-compensation'). In universities, poor teaching may often be masked by this, and final performance relatively little affected. On this view, studying the effect of courseware in isolation is unrealistic, but overall performance depends mainly on the

students' self-management rather than on any one resource. One could, however, attempt to study which resources students use and value (Brown *et al*, 1996).

6. Halo effects can also be important, where a teacher's attitude to the technology may strongly affect students in either positive or negative directions. While we have observed marked effects of this kind on student attitudes, whether this matters for learning depends also on whether student attitudes affect their use of the CAL software; if they have no alternative resource, it may not matter.

7. Similarly, Hawthorne effects may occur, where the act of doing the evaluation may affect students by making them feel more valued, pay more attention to the CAL software, and perhaps the subject matter it deals with. In addition to an effect on learners' attitudes, pre-tests given as part of an evaluation may well improve learning by communicating to the students what they should try to learn from the material. Furthermore, priming students to activate the relevant part of their theoretical knowledge is known to have a big effect on improving learning outcomes from laboratory classes and simulations, whether this is done by the evaluation or by a deliberate part of the teaching (for example, pre-laboratory exercises). All one can evaluate is the combined effect of the CAL software, the evaluation, and the whole of the associated teaching and learning resources. This is fine from the point of view of improving learning and teaching, and indeed evaluation should probably be a permanent part of practice, but it again undermines the view that the effects of CAL can be studied as an independent topic.

CAL is only part of an ensemble

The fundamental point is that CAL does not cause learning, and in fact is not a major cause at all. Learning results from the combined effect of many important factors, and typically, in universities, from multiple resources. Any realistic study or evaluation measures the combined effect of an ensemble. This does not mean evaluation studies are impossible, but it does mean we need to think out what we really want to discover. We cannot expect to treat CAL like a washing machine: as a simple device whose performance can be measured once in a standardized situation which will then tell us all we need to know to decide whether and how to use it.

If we remember that testing a piece of CAL is essentially the same as testing a textbook would be, then this seems obvious. It is also like considering the question: Is the 9.30 Glasgow–Edinburgh train good for getting to Edinburgh? It is possible to imagine that there could be something uniquely good or bad about that train and not others, but in fact usually the important factors are not the details of the train itself but how it fits into people's overall travel needs and plans. People use trains only as part of wider plans, and trains are mainly good or bad to the extent that they fit, or do not, into the success of these wider plans. To do meaningful evaluation of CAL, we have to understand learners' wider plans and study them: what they are, what the main factors are that influence their success, and where CAL fits into this.

The issues listed above are the symptoms of the basic theoretical problem with summative evaluation, which is that CAL does not cause learning like turning on a tap, any more than a book does. Instead it is one rather small factor in a complex situation.

Integrative evaluation: the actual utility of summative evaluation

In the TILT project, we performed many evaluations on completed software in the classroom. We found that our evaluation reports were often useful to teachers, but not for summing up the properties of the software so much as for identifying specific problems in the case being studied that the teachers would use to make changes, usually not to the software but to some other aspect of the delivery, for example, modifying how they introduced the software, or adding a section to a lecture. We thus realized that the value of our evaluation was not as summative evaluation of the software, but as formative evaluation of the overall teaching and learning situation. We called this 'integrative evaluation' (Draper *et al*, 1996) because most of the changes made concerned improving the embedding or integration of the software into the rest of the surrounding delivery.

Experience elsewhere seems to bear this out: often when CAL is first introduced there are some problems, particularly if students feel they have been set adrift without the right kind of support and guidance. If evaluation is taken as measuring how good or bad the software is, it would have to record a partial failure, and would have to refrain from making contributions to improvements. However, this is not how sensible participants actually use it: they use the evaluation to alert them to problems and quickly introduce improvements. The next time the course is given, the evaluation can often verify the improvement, but its most important role has been identifying problems and necessary improvements. Conversely, when participants treat an evaluation as strictly summative, problems can result. In an unpublished case, an evaluation was commissioned to help the institution decide whether to adopt a substantial CAL package as part of a course. The report was then interpreted strictly as supporting the adoption, but all its advice on the crucial integration issues it identified was overlooked and no further evaluation was permitted. Standard student feedback later indicated substantial quality problems on the resulting course, apparently around those integration issues identified in the original study.

What we really want, and what we can do

Evaluation is worth doing only if it serves some purpose and leads to some useful action. The previous section described how evaluations on completed software are often used to identify integration issues and so improve the total teaching and learning. Are there, then, any other goals that a summative evaluation might be required to help us with? Yes – the goals informally expressed as: Are we going to use it? How are we going to use it? An important decision that should be supported by information from evaluations, in education as in consumer purchases, is whether or not to buy and adopt some product, in this case a piece of CAL software. Relevant to this would be answers to whether students do learn in courses adopting the CAL software, whether this worked in other institutions (particularly one's own), what it costs in resources to run the course with the CAL software, and what one needs (and needs to know) in order to run such a course successfully: a description of the whole teaching delivery including auxiliary materials would thus be very desirable.

The fact that learning outcomes depend on the combination of many factors, of which the CAL software is only one, means that no single study can prove that the software will

work in any other situation; for example, it might work for the authors, but not work in the different context of a new adopter. However, while this does mean that certainty is beyond reach, it does not mean that evaluation is worthless for this. Imagine what you would find useful and persuasive as evidence about whether to adopt a piece of CAL software. The first important thing to discover is whether it has ever been used successfully, in other words with satisfactory learning outcomes. Even in fields much better understood theoretically, such as building aircraft, the first use is crucial to demonstrate that no crucial mistake has been made: one does not expect to be the first person on an aircraft never used before. But unlike aircraft, the performance of CAL depends to a great extent on the surrounding context of use, so tests with real students as part of a real course are much more convincing than tests in a laboratory with paid subjects. Thus, even though certainty may be beyond reach, tests showing that the CAL software can be successful are an important reason for and outcome of summative evaluation. Beyond that, the questions (see above) shade into issues of how best to use it. The more these issues are identified explicitly and made available in reports or auxiliary material for teachers, the better for those deciding whether to adopt it. Pure demonstrations of possible success, and outcomes from integrative evaluation, can together serve the essential underlying goal of supporting decisions about whether to adopt the software.

Experiments

A related issue is the use of controlled experiments. Summative evaluations on consumer goods are based on such controlled experiments, for example using the same standard load of dirty clothes and the same detergent for each washing machine compared. On the other hand, the useful results of integrative evaluation are often (though not always) the result of open-ended observation or student feedback, identifying factors that had not been foreseen and so not systematically measured. Furthermore, all the points above about the factors likely to be important in affecting outcomes suggest that we do not know enough to control all the relevant factors. Few if any experiments, for instance, attempt to control the halo effect or students' uncertainty about study strategies with CAL.

Can experiments be used meaningfully at all in CAL evaluation? This remains arguable. Clark (1983) has contended that no meaningful experiments on whether learning is affected by the medium of instruction have been or could be done, because other factors more plausible as the causative agent vary with the medium. While hotly debated in the literature (Ross, 1994), his arguments have not been conclusively rebutted, and they apply also to the use of experiments in evaluation. They would apply most strongly to prevent us from drawing generalizations such as that a piece of CAL software will always be successful. Nevertheless, some experiments seem rather convincing, for example, MacDonald and Shields (1996), particularly when alternative ways of teaching are directly compared, for example lectures, and CAL with and without special worksheets. Note that such experiments can in part serve an integrative role by yielding information on how best to use the software: there is no exclusive association between evaluation methods (experiment versus open-ended observation) and the evaluation goal (summative or integrative). In the end, experiments are probably like other studies of CAL: they can

show that the CAL software was definitely part of successful learning in one case, and if no other cases have been reported, then this is in favour of the software. On the other hand, we remain aware that many factors may have been important, and experimental reports in the literature never describe them all. It could always be that one of these factors was crucial, but will not be present if one tries to use CAL in one's own teaching; for instance, freeing up lecturers (by substituting CAL for lectures) perhaps meant that while in the laboratory supervising CAL use, they performed tutorial interactions with students that were crucial, just to have something to do.

Conclusion

We can do evaluation that is summative in some senses but not in others. We can do evaluation at the end of the design cycle on completed software that will not be further modified. We can do evaluation that provides evidence relevant both to deciding whether to use that piece of courseware for teaching, and on how best to use it. This latter evidence might be from careful comparative experiments or more formative style work which detects unforeseen problems that turn out to be important in successful use of the software. But there is no prospect of doing evaluation that sums up a product once and for all, measuring its essential properties in a way that will represent and predict its performance in all other contexts.

Acknowledgements

This paper stems from work on the TILT (Teaching with Independent Learning Technologies) project, funded through the TLTP (Teaching and Learning Technology Programme) by the UK university funding bodies (DENI, HEFCE, HEFCW, SHEFC) and by the University of Glasgow. The ideas come from collaboration with other members of the evaluation group, particularly Margaret Brown and Erica McAteer. The studies mentioned here could not have been done without, in addition, the active participation of many members of the University teaching staff to whom I am grateful.

References

Brown, M.I., Doughty, G.F., Draper, S.W., Henderson, F.P. and McAteer, E. (1996), 'Measuring learning resource use', *Computers and Education*, 27, 103–13.

Clark, R.E. (1983), 'Reconsidering research on learning from media', *Review of Educational Research*, 53 (4) 445–59.

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. and Watt, H.D. (1995), *Using Learning Technologies: Interim Conclusions from the TILT Project*, TILT Project Report no.3, Robert Clark Centre, University of Glasgow.

Draper, S.W., Brown, M.I., Edgerton, E., Henderson, F.P., McAteer, E., Smith, E.D.

and Watt, H.D. (1994), *Observing and Measuring the Performance of Educational Technology*, TILT Project Report no.1, Robert Clark Centre, University of Glasgow.

Draper, S.W., Brown, M.I., Henderson, F.P. and McAteer, E. (1996), 'Integrative evaluation: an emerging role for classroom studies of CAL', *Computers and Education*, 26 (1–3), 17–32; and in Kibby, M.R. and Hartley, J.R. (eds.), *Computer Assisted Learning: Selected Contributions from the CAL 95 Symposium*, Oxford: Pergamon, 17–32.

MacDonald, Z. and Shields, M. (1996), *Effective Computer-Based Learning of Introductory Economics – Some Results from an Evaluation of the Winecon Package*, Discussion Paper in Economics No. 96/11, University of Leicester.

Ross, S.M. (1994), 'Delivery trucks or groceries? More food for thought on whether media (will, may, can't) influence learning', Introduction to special issue of *Educational Technology Research and Development*, 42 (2) 5–6.