Evaluating online teaching and learning

Martin Oliver
University College London, Gower Street, London, WC1E 6BT
Tel: +44 20 7679 1905; Fax: +44 20 7679 1715; E-Mail: Martin.Oliver@ucl.ac.uk

Abstract. Evaluation is becoming increasingly important, both as a part of the design of online courses and as a mechanism for quality assurance. In this paper, the issues facing the evaluation of online teaching and learning are considered. Different motivations for evaluation are identified, and strategies for addressing these needs are illustrated.

Introduction

The drive for transparancy and public accountability in the UK's public sector has had a far-reaching impact on Higher Education. Part of this impact has been an increased emphasis on evaluation (Oliver, 2000). However, the drive to evaluate has not been matched by support and training for the practitioners who are supposed to carry out these processes (see, e.g., Phelps et al., 1999).

In response to this, several initiatives have been implemented to provide practitioners with support, such as the development of toolkits (Conole et al., 2000), cookbooks (Harvey, 1998) or manuals of advice and guidance (Phillips et al., 1999). What these resources lack, however, is specific advice on evaluating online learning and teaching. Consequently, this article will include a review of the issues specific to this domain, supplemented by illustrative cases. To structure this, however, it is necessary to elaborate the reasons for evaluating online learning and teaching.

Background

The characteristics of distance learning and of online learning and teaching

As noted above, many discussions in the literature address generic issues of evaluating learning technology rather than concentrating on the particular characteristics of online learning and teaching. However, it is important to take these into account when designing and implementing an evaluation. In order to do this, the charactersitics of distance learning will be outlined, and then extended by a consideration of online learning and teaching.

Peters (1998) identifies several distinctions betwen distance learning and traditional forms of study. These include:

- A shift from an elitist model to mass higher education
- A move towards increasingly structured and planned programmes of study
- The industrialisation of course development, including the division of labour amongst teams of specialists
- The challenge of maintaining dialogue as a central component of distance courses

- The loss of informal opportunities for learning, for example in social settings

Importantly, although physical distance is taken into account in his analysis, Peters is primarily concerned with "distance pedagogics". Such techniques as broadcast lectures, because they add nothing to traditional forms of teaching, are grouped with traditional teaching techniques. By contrast, open learning self-instructional materials are considered alongside distance learning techniques, even when the specific packs are distributed to campus-based students. In this paper, however, physical distance will be considered, since it introduces a range of pragmatic problems to evaluation.

Online learning and teaching is harder to characterise, due to its relatively short history and diverse forms of implementation. Clearly, most online programmes will mirror the characterisation of distance learning provided above. Importantly, however, several new characteristics may also be added. These include:

- Technical requirements
- Skills requirements
- The breadth and use of different media

These characteristics are equally important for staff and students involved in online learning. (See, e.g., Salmon, 2000.)

Another important aspect of online learning and teaching is that many systems automatically log use. Particularly important is the way in which online discussion – compared closely to traditional correspondence learning by Peters (1998) – is recorded in full. As will be seen, this is of considerable use to evaluators.

Evaluating learning technology

The term 'evaluation' refers to a wide-ranging collection of methodologies, and is also used to cover review processes such as checklists as well as empirical judgements. It is worth noting that the term is sometimes confused with assessment (Phillips et al., 1999); however, in this article, it is taken to cover processes that support judgements of value and worth of programmes (Guba & Lincoln, 1981).

Reviews of evaluation methodologies have stressed the importance of determining the *purpose* that the process will serve. Numerous distinctions have been made, but the following set (from Oliver, 1997) has been adopted as a useful summary within the learning technology community:

- Formative evaluation
- Summative evaluation
- Illuminative evaluation
- Integrative evaluation
- Evaluation for Quality Assurance

These five purposes will be adopted to structure the following discussion. In addition, special attention has been given in the literature to the evaluation of costs and to comparative evaluations. As they represent special cases, these two categories will also be considered.

Summary

In this section, the scope of this review has been established. The key characterstics of online learning and teaching and reasons for evaluation have been identified. In the next section, the evaluation of online learning and teaching will be considered, starting with general issues before moving on to consider each of the specific reasons for evaluating a programme in turn.

Evaluating online learning and teaching

General issues for evaluation

In the previous section, the characteristics of online learning and teaching were identified. Of all of these, the most immediate impact on evaluation arises as a consequence of the physical distance involved. Even a cursory glance at lists of methods for gathering data (e.g. Oliver & Conole, 1998) will reveal that most involve contact with the students. Unless considerable effort and expense can be made to arrange meetings, methods such as focus groups, interviews and observation are rendered impractical.

Many suggestions have been made that re-create these methods by proxy. Cousin & Deepwell (1998), for example, have discussed the feasibility of virtual focus groups. They demonstrated that these can be an effective substitute for a real meeting, and offer all the benefits often advocated for computer-mediated communication such as allowing space for reflection when responding. However, they also noted several limitations, such as participants' reluctance to contribute messages on sensitive topics and the need for a skilled facilitator of online discussions (see, e.g., Salmon, 2000).

An alternative approach is to focus the evaluation on the types of data that online systems are good at gathering. Phelps & Reynolds (1998), for example, combined web-based questionnaires (allowing immediate responses without the subsequent need for lengthy data entry) with system usage data such as the time and frequency of page access. This was achieved by using Javascript to create a tracking log via a CGI script on the host server each time a page was requested. These method provided very rich data on patterns of usage and on users' motivation and satisfaction. However, once again, the methods had their limitations. Usage logs are difficult to interpret, since they cannot reveal why a learner accessed a particular page or what they did with it once they had gained access. Whilst this type of data is valuable and easy to collect, it remains important to triangulate it with other sources as part of the interpretative process.

Additionally, in this case, the rate of return for the online questionnaire was low. This leaves the evaluation open to the criticism that the opinions recorded will be from a self-selecting sample, and thus unrepresentative of the wider group of users. In particular, it seems probable that less confident users of technology are those least

likely to respond. Other evaluations of online learning (e.g. Taylor et al., 2000) have complemented this method using a paper-based survey distributed to non-respondents of the online questionnaire, leading to a much better overall response rate.

Another common approach to data gathering in an online environment involves the creation of a "feedback" discussion area. Again, this offers the opportunity for continuous feedback from participants and also provides a full transcript of responses in an electronic format, ready for analysis. However, as Taylor et al. (2000) note, this can open the floodgates to an unstructured wash of criticism, much of which may come from a small but vocal minority. These views, which may be unrepresentative, can cause considerable problems if used during formative assessment if designers feel that they need to take all criticisms into account. Since it is impossible to please all of the learners all of the time, the value of feedback forums may be greatest when an evaluator is able to act as an intermediary between the data and the course team.

Many of these issues can be summarised by noting that evaluation in this context raises two general problems for evaluators. Firstly, with many of these methods, the process can no longer be controlled. Opportunities for contributing data can be provided, but what the student does with this opportunity is up to them. Taking a more extreme position, it is also impossible to tell who is actually contributing the data. Secondly, methods for interpreting these types of data are still being developed. Whether the data be from system logs or bulletin boards, lessons are still being learnt about the most useful and appropriate ways of drawing conclusions.

Finally, it is worth identifying methods of evaluation that are *not* affected by the move to distance education. Essentially, these will be either those designed to operate at a distance, such as postal or telephone surveys, or those that do not require empirical data. Surveys are clearly subject to the same issues as other distance methods, as discussed above. Methods such as checklists and conceptual maps, which fall into the latter category, are also subject to criticism. Whilst these are relatively easy to implement, significant questions have been raised about their value (e.g. Tergan, 1998), not least because of their highly subjective nature.

Formative evaluation

Having highlighted some pragmatic problems for the evaluation of online learning, it is worth considering the purposes of evaluation. In the previous section, five purposes were identified; the first of these is formative evaluation. This refers to evaluations that are intended to provide information that allows revisions and improvements to be made. Its primary audience usually consists of the project or course team.

Several features characterise formative evaluations. Firstly, they are usually carried out by a member of the project or course team; in this regard, they are 'internal' evaluations. In order to be useful, they must provide timely information in a format that is readily accessible to the course team. In this respect, utility is a higher priority than validity (Patton, 1997).

Within small, self-contained teams, immediately accessible evaluations (where the results need little or no subsequent analysis) such as focus groups are often useful. As noted above, however, such techniques are often inapplicable in a distance context. In addition, any source of data that relies on an input from the students will introduce

delays into the evaluation when nothing can be done but to wait for responses. The implication of this is that scheduling becomes extremely important. Ironically, although formative evaluations may be what is referred to as "quick and dirty", they are at their most useful when carefully prepared for. The economy of effort must come in the collection and analysis of data, rather than in the planning of the study.

A good illustration of the issues involved in formative evaluation is provided by the EuroMET project (Phelps & Reynolds, 1999). This involved the development and delivery of web-based courses in Meteorology by a consortium of 22 partners. Given the complex structure of the project, it was important to ensure that appropriate information was gathered and communicated in a timely manner. The evaluation that was carried out included two strands. The first was a survey of users' views on ease of use, pedagogy (including scientific integrity) and value as a replacement for traditional teaching methods. The second involved usability trials, carried out with a sample of five users under controlled conditions. This approach allowed the project team to identify elements of the course that worked and those that needed revision. The strong use of visual material was welcomed by users, for example, whilst inconsistencies in the material (such as variations in style and symbol use from section to section, reflecting the different contributing authors) were identified as an area for attention. Both strands of evaluation contributed to the re-design of the system's navigation. The evaluation showed that users found that some icons were too similar to each other, that users had no sense of where they were in the material, and that students wanted clear learning objectives and end-of-unit summaries to be added.

The timing of this evaluation allowed these points to be fed back to the project team and suitable revisions incorporated. The structure had added value in that one strand helped to validate the revisions that were proposed.

> *The formative evaluation has been extremely useful in producing modules which are suitable for their target audience, are easy to use, and are instructive. The fact that the evaluation was embedded into the development work meant that it was relatively easy for the developers to modify the modules according to the recommendations of the evaluators and, in turn, for these modifications to be tested during the next evaluation phase. In particular the usability study showed that the modifications made after the first evaluation phase were effective.*
>
> *(Phelps & Reynolds, 1999, p. 192.)*

Summative evaluation

In contrast to formative evaluation, summative evaluation is often an external process concerned with judgement rather than improvement. It often involves assessment of a project against its aims or, in the case of online education, of a course in terms of learning outcomes. It is often asserted that such evaluation ought to be carried out by an evaluator outside of the project team in order to assure objectivity (e.g. Bradbeer, 1999).

However, recent critiques of evaluation have made the point that evaluation is inherently political (Patton, 1997); objectivity is, in many ways, a myth. Many of the proposed advantages of scientific methods, designed for use in controlled conditions,

such as transferability and replicability, simply do not apply in the ill-defined, authentic world of education practice.

Such critiques have led to a division between experimental designs for summative evaluation and those that are primarily exploratory (Oliver & Conole, 1998). This section will focus on experimental approaches; the exploratory approaches will be explored further in the following two sections.

Experimental approaches face several challenges. One of the most significant is that it is effectively impossible to prevent 'contamination', where some factor external to the experiment influences outcomes. An obvious example of this would be if an online course broke down, and students passed the final exam because they had all formed self-help groups and taught themselves from textbooks instead. Since it is impractical to control all the extraneous factors in any educational setting, particularly when it involves learning at a distance, it becomes extremely difficult to attribute causality to the teaching intervention.

More subtle problems arise in the context of comparative evaluations. These are often popular with managers or funders, since it is assumed that the comparison will demonstrate whether the innovation adds any value to the learning experience. Here, however, cross-condition contamination is even harder to prevent. Even if online courses are password-protected, it is quite possible for students to share IDs, download materials or even just share notes.

Other problems also arise for comparative experiments. Experiments are predicated on the ability to control the context in which they take place; this is necessary in order to isolate the variables to be studied. In an educational setting, it is often impossible to do this on pragmatic and ethical grounds. If different teachers are involved, another important factor is introduced. The same is true if the materials change, the students are different, the subjects covered vary or the way in which they are taught alters. From an ethical point of view, it is difficult to justify allocating extra resources or opportunities to only a sub-set of a student body, particularly when the course carries credit towards an award.

A final criticism is aimed at comparisons of traditional and computer-based courses. This is particularly relevant for courses that are subjected to a comparison of learning outcomes "before and after" adaptation to an online format. The argument is that, because the methods used differ so radically, these experiences are so different that they cannot be compared in any meaningful way. The analogy used is that it is like comparing apples and oranges.

It has been argued that such comparisons can be drawn, but that this must be done with care (Oliver & Conole, 1998a). It is an easy matter to gather data on student preferences, for example, or to compare performance on an end of year exam. What must be asked, however, is what such a comparison *means*. If the change from traditional to distance learning (for example) really does represent a completely new educational experience, it is inappropriate for the assessment used to remain the same - a point often neglected when designing online courses. This raises serious questions about the validity of assessment methods which is beyond the scope of this paper. However, if the assessment does remain the same, then a comparison of performance - irrespective of the measure's validity - can clearly be made. If what is being

evaluated is simply student performance against some assessment yardstick, then it is appropriate to compare their net experience in reaching this. In such cases, comparison of courses becomes a sensible option.

Given the problems noted above about contamination, control and transfer, what the experimental approach may permit is a firm conclusion about one particular comparison (albeit with the proviso that the measurement's reliability should be considered critically). Claims about transferability, however, are more difficult to justify. The implications of this are that experimental evaluations (and comparative evaluations) are possible. However, they must be designed with care, reported in a way that acknowledges the limitations of the method in an educational context, and interpreted with the same criteria as any qualitative case study.

Hiltz et al. (2000) provide a good example of an evaluation of online course that adopts a critical approach to experimental methods. The evaluation concentrates on the Virtual Classroom® system, and involved three separate studies. These considered hypotheses such as, "[Online communication and learning] can improve quality of learning as measured by grades or similar assessments of quality of student mastery of course material". Importantly, the proviso made above about the validity of assessment as a measure of learning is explicitly acknowledged here.

Careful attention is paid to general experimental evaluation issues, and explicitly discusses the limitations of the experimental method outside of a laboratory setting. Moreover, the limitations of studies are also made clear. For example, considering the first study of Asynchronous Learning Networks (ALNs) described in isolation, Hiltz et al. acknowledge that, "The longitudinal field study does not allow us to conclude whether better educational outcomes in ALN-supported courses are the result of collaborative learning techniques, ALN use, or both." This problem was tackled by triangulating the three studies. This allows Hiltz et al. to conclude that, "though any one measure or method might be legitimately questioned in terms of its validity, reliability, or generalizability, the weight of several different kinds of studies over a period of five years, is convincing."

Illuminative evaluation

The problems of employing experimental methods in educational settings are not new. In 1987, Parlett & Hamilton proposed an alternative model based on a 'social anthropological' approach to evaluation. Rather than attempting to quantify impact, these studies seek to discover the factors that are important to the participants. This is achieved through phases of observation, inquiry and explanation, with analytical methods adapted pragmatically and triangulation used to improve the reliability of findings.

In contrast to experimental evaluations, which seek to control the factors that might influence learning and teaching, illuminative evaluation seeks to describe and interpret them. The educational context becomes the focus of the study, rather than the measure of learning that is used for assessment purposes.

This freedom clearly avoids the problems that faced experimental approaches. However, the very responsive flexibility that allows illuminative evaluation to achieve this prevents its conclusions from being objective or transferable. Conclusions are

interpretations constructed by the evaluators. Confidence in them can be increased if methods are triangulated, but they remain interpretations of specific events. Although this may be perfectly adequate for summative evaluations of single programmes, it will pose problems if the lessons learnt from this are to be applied elsewhere. Experimental studies make the claim that their results are generalisable; however, as noted above, this claim is problematic in an educational setting.

In summary, illuminative evaluation accepts the criticisms levelled at experimental studies and, rather than trying to overcome them, works within the constraints that they represent. No attempt is made to generalise, for example. ***Work on this!!!

Wegerif's study of the development of communities in online discussion provides an example of the illuminative approach to evaluation. The study involved participant observation, in-depth interview together with surveys and a transcript of discussion, then analysis. This allowed a deeper understanding of the process through which students succeeded or failed in joining an online community, and the implications of this on their achievements. The conclusions that were drawn were specific to this situation, but recommendations were put forward as a starting point for discussion – including comparison with other studies of situations like this.

> *"As well as its more specific conclusions and recommendations, this study has illustrated a method for researching the social dimension of ALNs and put forward the beginnings of a conceptual framework, including the concept of the difficult threshold between insider and outsider status, which may prove of general value in understanding the impact of the social dimension on learning on ALNs and how this impact can be taken into account in course design."*
>
> *Wegerif, 1998*

Integrative evaluation

Experimental and illuminative approaches to evaluation can be seen as two extremes, each of which has limitations. Several evaluators have attempted to create compromises that incorporate elements of both approaches. Integrative evaluation is one such approach.

The term integrative evaluation is used in several contexts, but in the field of learning technology research is usually associated with the approach devised by the Teaching with Independent Learning Technologies (TILT) project (Draper et al., 1994). This combined the structured approach of experimental evaluations with the values and flexibility of illuminative studies. In addition to the study itself, phases of work took place that addressed the context of the course, addressing issues such as policy, resources and the tacit teaching objectives of the staff involved. Integrative studies incorporate multiple methods, including within-group experimental studies of performance, surveys, interviews and confidence logs.

Inherent in the approach, however, is the assumption that the evaluation's findings will be situationally-specific. The term "integrative" reveals the central motivation for the project team, which was to improve the way that computer-based resources

(including online materials) were incorporated into the course. As with illuminative approaches, no attempt is made to generalise the findings.

Draper & Brown (1998) used the integrative approach in their study of remote collaborative tutorial teaching. This involved around 20 different studies, each of which adopted a similar approach, and which were then summarised and synthesised in order to make claims about the project as a whole. This process allowed the evaluators to argue that the collaborative tutorial teaching process were at least as effective as traditional methods, were received with mixed levels of enthusiasm, and were primarily of benefit in enriching the curriculum and in staff development. There was no need to generalise the conclusions beyond this point, since the approach was unique to this project; however, the synthesis of so many individual integrative studies did provide an adequate basis for summative judgement and for recommendations for others attempting to adopt a similar approach.

Evaluation for Quality Assurance

References

Draper, S. & Brown, M. (1998) Evaluating remote collaborative tutorial teaching in MANTCHI. In Oliver, M. (Ed) *Innovation in the Evaluation of Learning Technology*, 65-86. London: University of North London Press.

Draper, S., Brown, M., Edgerton, E., Henderson, F., McAteer, E, Smith, E. & Watt, H. (1994) *Observing and measuring the performance of educational technology*. TILT project report, University of Glasgow.

Wegerif, R. (1998) The Social Dimension of Asynchronous Learning Networks. *Journal of Asynchronous Learning Networks*, 2 (1). http://www.aln.org/alnweb/journal/vol2_issue1/wegerif.htm

Parlett, M. & Hamilton, D. (1972) Evaluation as Illumination: a new approach to the study of innovatory programmes. In Murphy, R. & Torrance, H. (Eds) *Evaluating Education: Issues and Methods*. London: Harper & Row.

Hiltz, S., Coppola, N., Rotter, N., Turoff, M. & Benbunan-Fich, R. (2000) Measuring the Importance of Collaborative Learning for the Effectiveness of ALN: A Multi-Measure, Multi-Method Approach. The Journal of Asynchronous Learning Networks, 4 (2). http://www.aln.org/alnweb/journal/Vol4_issue2/le/hiltz/le-hiltz.htm

Patton, M. (1997) *Utilization-focused evaluation*. London: Sage.

Taylor, J., Woodman, M., Sumner, T. & Blake, C. (2000) Peering Through a Glass Darkly: Integrative evaluation of an on-line course. Educational Technology & Society, 3 (4). http://ifets.ieee.org/periodical/vol_4_2000/v_4_2000.html

Phelps, J. & Reynolds, R. (1999) Formative Evaluation of a Web-based Course in Meteorology. *Computers & Education*, 32, 181-193.

Phelps, J. & Reynolds, R. (1998) Summative Evaluation of a Web-based Course in Meteorology. In Oliver, M. (Ed) *Innovation in the Evaluation of Learning Technology*, 135-150. London: University of North London Press.

Salmon, G. (2000) *E-Moderating: The Key to Teaching and Learning Online*. London: Kogan Page.

Cousin, G. & Deepwell, F. (1998) Virtual Focus Group Techniques in the Evaluation of an Electronic Learning Environment. In Oliver, M. (Ed) *ELT 98: Innovation in the Evaluation of Learning Technology Conference Proceedings*, 4-7. http://www.unl.ac.uk/tltc/elt/elt98.pdf

Oliver, M. (2000) An introduction to the evaluation of learning technology. *Educational Technology & Society*, 3 (4). http://ifets.ieee.org/periodical/vol_4_2000/v_4_2000.html

Phelps, J., Oliver, M., Bailey, P. & Jenkins, A. (1999) The Development of a Generic Framework for Accrediting Professional Development in C&IT. EFFECTS report no. 2, University of North London.

Conole, G., Oliver, M. & Harvey, J. (2000) An Evaluation Toolkit for practitioners: Scoping Study. JISC 'An Evaluation Toolkit for Practitioners' project report no. 1, University of Bristol.

Harvey, J. (1998) *The LTDI Evaluation Cookbook*. Glasgow: Learning Technology Dissemination Initiative.

Phillips, R., Bain, J., McNaught, C., Rice, M. & Tripp, D. (1999) Handbook for Learning-Centred Evaluation of Computer-facilitated Projects in Higher Education. http://cleo.murdoch.edu.au/projects/cutsd99/

Guba, E. & Lincoln, Y. (1981) *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*, London: Jossey-Bass.

Tergan, S. (1998) Checklists for the Evaluation of Educational Software: Critical Review and Prospects. *Innovations in Education and Training International*, 35 (1), 9-20.

Oliver, M. (1997) *A framework for evaluating the use of learning technology*. ELT report no. 1, University of North London. http://www.unl.ac.uk/tltc/elt

Peters, O. (1998) *Learning and Teaching in Distance Education: Analyses and Interpretations for an International Perspective*. London: Kogan Page.

Salmon, G. (2000) *E-Moderating: The Key to Teaching and Learning Online*. London: Kogan Page.

Oliver, M. & Conole, G. (1998) Evaluating Communication and Information Technologies: a toolkit for practitioners. *Active Learning* 8, 3-8.

Bradbeer, J. (1999). *Evaluation*. Technical report no. 8, University of Portsmouth.

Oliver, M. & Conole, G. (1998a) The Evaluation of Learning Technology - an Overview. In Oliver, M. (Ed) *Innovation in the Evaluation of Learning Technology*, 5-22. London: University of North London Press.