Contexts for evaluating educational software

A. Jones\textsuperscript{a,\*}, E. Scanlon\textsuperscript{a}, C. Tosunoglu\textsuperscript{a}, E. Morris\textsuperscript{a}, S. Ross\textsuperscript{b}, P. Butcher\textsuperscript{c}, J. Greenberg\textsuperscript{c}

\textsuperscript{a}Institute of Educational Technology, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
\textsuperscript{b}Open University Cambridge Regional Centre, 12 Hills Rd, Cambridge CB2 1PF, UK
\textsuperscript{c}Academic Computing Service, Milton Keynes MK7 6AA, UK

Abstract

The evaluation of educational software is of concern to two particular academic communities: HCI and educational technology. There is a danger that usability features are considered at the expense of educational issues (and the converse of this is of course equally true). This paper considers how the notion and practice of evaluation in the educational community differs from that in HCI and also identifies areas of overlap. It then describes how particular influences and contexts have led one group of evaluators from the educational community to develop a context, interactions, attitudes and outcomes (CIAO!) model of evaluation for computer assisted learning (CAL) evaluation in distance education. The application of this model is illustrated by case studies from a recent evaluation project and related research. The paper concludes with a discussion of the issues raised for both communities by this model for evaluation. © 1999 Published by Elsevier Science B.V. All rights reserved.

Keywords: Educational software evaluation; Contextual evaluation; CIAO! evaluation framework

1. An evaluation framework for educational software

Evaluation methods and curricula has become very important in a climate where there is increasing concern with assessing and maintaining quality. In the UK, Higher Education has seen a rapid expansion over the past decade, which has not been accompanied by an equivalent increase in funding. It has also, of course, been a period of rapid developments in computer technology and growth in its use. In their search for new ways of teaching large numbers of students, universities have considered how the use of new information and communication technologies might help. Many applications not only involve new technologies but offer opportunities for teaching and learning in new ways. For example, much multimedia involves the increased use of resource based learning, whilst various
forms of electronic communication facilitate the development of learning communities and peer learning. Given the importance of determining the effectiveness of the uses of information and communication technologies for learning, the need for evaluation becomes crucial.

This paper considers evaluation of educational software from the perspective of an educational setting. However, educational software evaluation is also commonly undertaken by the HCI (human–computer interaction) community and we discuss the differences (and common ground) between the two contexts: first by looking briefly at usability and learning and then by considering the use of a particular framework for CAL evaluation which is illustrated by case study evaluations.

1.1. Usability and learning

Squires and Preece [1] argue that “there is no consideration of the implications of usability features for the use of the package to achieve educational goals (in evaluating educational software). There is a need to help evaluators consider the way in which usability and learning interact.” In the Jigsaw evaluation model proposed by Squires and Preece [1], and in the evaluation framework advocated by Squires and McDougall [2], the focus of attention is the predictive evaluation of educational software: “i.e. the evaluation of software prior to its use, which, typically occurs when teachers are either planning lessons or making purchasing decisions” [1]. They argue that thinking of learning and usability as independent issues leads to superficial evaluations of educational software and that many teachers are not trained to consider usability. This argument is supported by an analysis of two well known checklists (one developed by the National Council for Educational Technology for evaluating CD-ROMs and the American Microsft Evaluators Guide) which reveals the lack of attention to usability and the lack of integration between the two issues.

The Jigsaw evaluation model is designed to increase teachers’ awareness of the integration of learning and usability issues. In this model, the learning task consists of the specific task concepts needed for the topic and the general domain concepts. Similarly, the operational task includes interaction with the interface and also with the underlying operational system. These four tasks: specific learning tasks, general learning tasks, application operational tasks and system operational tasks are first of all considered independently at the first task level. At the second level, integration within the learning and operational tasks is considered and at the third level, integration between the learning and operational tasks. Two studies were conducted to investigate the support that the Jigsaw model could give teachers in selecting software for their students. In the first study, teachers chose a familiar educational application and in the second they were given the same package to evaluate. It was found that after using the Jigsaw model, teachers showed more awareness of integration.

The teachers ran the software themselves, but did not use it with students and in this context the model seemed to help the teachers’ understanding of how usability issues relate to learning outcomes. A further important issue, however, is the extent to which software can be properly evaluated without using it with students: it is in the context of the evaluation that the methods and practice of the different communities diverge. For many
evaluators, concerns include questions such as: ‘how successful was the software in achieving its intended aims and how was it received by the students?’ and so it is imperative to involve learners in the evaluation. Squires and Preece [1] discuss the possibility of applying their model to ‘interpreive evaluations’ i.e. evaluating the users use of the software, but their focus is on the predictive evaluation. We suggest that one reason for the gap in approaches, therefore, is that the goals are quite different: in the context Squires and Preece describe, the goal is selecting educational software and in ours, evaluating the usefulness and success of particular software.

The term ‘interpreive evaluation’ itself may be unfortunate as it suggests a contrast to the more ‘objective’ nature of predictive evaluation. Yet all evaluations, whether with learners or not, are located in particular contexts which makes the process one of interpretation. This was recognized many years ago in the approach taken by Parlett and Hamilton [3] who viewed their illuminative evaluation approach as belonging to “a(n)...‘anthropological’ research paradigm. Attempted measurement of ‘educational products’ is abandoned for intensive study of the programme as a whole: its rationale and evolution, its operations, achievements and difficulties. The innovation is not examined in isolation but in the school context or ‘learning milieu’”. This view was more recently supported by proponents of situated learning (e.g. Brown et al. [4]) who argue that the social and cultural contexts of learning are a crucial part of that learning.

In the next three sections we briefly outline the history that led to our particular evaluation framework. The framework is then introduced and illustrated through the discussion of three case studies. Finally we return to consider some of the main issues for both the HCI and educational communities in the evaluation of educational software.

1.2. Early evaluations at the OU

The evaluation studies described here were carried out at the Open University which has a long history of the use and evaluation of CAL. The first coordinated evaluation of the use of CAL, (which was introduced in the mid 1970s) was conducted in 1979 and focused on the science faculty. The main conclusion was that students had realistic expectations about the benefits that might be obtained from computer use, but were also faced with a number of difficulties which included problems of access, integration with other course material and lack of support. Affective factors were also important, students had three particular anxieties: fear that they might look stupid; fear that they might ‘break’ the software; and fear that they might be spied on, i.e. that the regular computerized assessments might be linked to their performance on remedial CAL tutorials [5].

In order to uncover these issues it was necessary to adopt a mix of qualitative and quantative methods and to adapt methods appropriately in order to discover unexpected as well as expected outcomes. For example, the students’ fears did not emerge from the first questionnaire study, but from informal interviews at residential schools. This supported Parlett and Hamilton’s [3] argument that an illuminative method is needed to uncover what’s happening, not just what the evaluators think is happening. This same approach had been adopted in the evaluation of a national programme, The National Development Programme on Computer Assisted Learning [6].
1.3. Iterative cycles of development and evaluation

Elsewhere one of the authors has discussed an iterative model of formative evaluation [7] which was applied by the Open University to the development of stand alone teacher training materials (including software). This included three evaluation phases. First, external readers provided critical comments (in addition to internal comments). In the second phase students worked through the whole course in detail. Finally, because the integration of practical computer based work with the text was crucial, the revised version was again tested out by a small group of students.

The need for this level of formative evaluation is telling. The university invests heavily in the design and development of its courses and the critical commenting process is extremely helpful in identifying problem areas. Yet the author’s conclusion was that the only way to know how learners will use a particular course and what problems they experience is to study them using it. Finding out what happens in practice with real students is crucial: “there is no design process or method that can predict the outcomes or pre-empt all the learner’s problems. So the design process must be iterative: the material must be tried out on students, and refined, and then tried again, and the cycle continued for as long as necessary” (Jones et al. [7] pp. 119).

An additional issue is the constructivist nature of learning: “learners are active and interpret what they learn in the light of the knowledge they already have. We can therefore never assume that a piece of instructional material, studied by a learner at a distance, is being used or interpreted in the way that the designer intended or hoped” (Jones et al. [7] pp. 119). This strong belief in the need to involve learners and to try to understand their perspective, is an important factor in influencing the OU’s approach to CAL evaluation.

1.4. Evaluations in the 1980s and 1990s

During the 1980s the university, like many other institutions, experienced a growth of interest in two kinds of collaborative use of computers: Computer Mediated Communication (CMC), which was experimented with on a small scale and the use of computers by groups of learners.

One study by Issroff [8] compared the performance of individuals and pairs in the collaborative use of CAL in chemistry. Like many other evaluation studies there were no differences between the individuals and pairs when learning outcomes were measured by post tests. However, the perceptions and motivations of the students in the two conditions differed: the students who worked in pairs were more positive, enjoyed the experience more and were more motivated. Differences were also apparent in the work produced by the students whilst they used the computer (in this case worksheets) and social factors affected the nature of the interaction. This study confirmed the importance of paying attention to affective aspects as well as cognitive and to using a variety of measures which captured students’ interactions and work whilst using the program as well as measures of learning outcomes.

The early CAL evaluations and later large scale evaluations of computer use, studies of collaborative use of computers and formative course evaluations were all important factors in influencing the model of evaluation developed for a new CAL evaluation project in
1994. The evaluation literature was also carefully considered [9]. One of the concerns emerging from this review is the effectiveness of CAL materials. For example, the goal of Muller’s evaluation methods [10] is to obtain empirical evidence of student performance in order to judge the instructional effectiveness of software. A further conclusion of Muller’s review [10] was that comparative approaches to evaluating effectiveness are problematic. Further, findings from comparative studies usually only yield information about what has occurred (e.g. whether students have learned more efficiently with CAL or not) and not how this learning has occurred. Yet often the ‘how’ is crucially important, for example, in order to help learners change their misconceptions and help correct their errors we need to know what the misconceptions are and what kinds of errors they make, not simply that they make them. This emphasis on the process of learning also emerged in the literature review [9].

2. The CIAO! framework for evaluation

The context, interaction, attitudes and outcomes (CIAO!) framework was developed to support the evaluation of CAL developments at the university. Many of the CAL packages are in science and some of these evaluations are reported elsewhere in Scanlon et al. [11]. In this section, the CIAO! framework is described and it is followed by case studies of how it worked in practice. The evaluation team included CAL designers, course developers and educational technologists and so had access to the stated aims and rationales of the CAL that was looked at and also knowledge about the context of use. Our concerns were to evaluate the effectiveness and quality of CAL whilst simultaneously investigating the educational situation as a whole and focusing on the learners. The approach is one in which information from a variety of sources and both qualitative and quantitative techniques are used. The framework proposed in order to do this [9] outlines three dimensions to evaluate: (i) context; (ii) interactions; and (iii) attitudes and outcomes. This is summarized in Table 1 and the rationale for the context, interactions, attitudes and outcomes cells is described later.

<table>
<thead>
<tr>
<th>Rationale</th>
<th>Context</th>
<th>Interactions</th>
<th>Attitudes and outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designers and course team’s aims. Policy documents and meeting records</td>
<td>The aims and the context of use</td>
<td>Data about students interactions with the software allows some analysis of the process.</td>
<td>It is important to try to assess learning outcomes but also to consider effective outcomes, e.g. perceptions and attitudes</td>
</tr>
<tr>
<td>Interview CAL designers and course team members. Analyse policy documents</td>
<td>Records of student interactions. Products of students work. Student diaries. On-line logs</td>
<td>Measures of learning. Changes in students attitudes and perceptions</td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td>Interview CAL designers and course team members. Analyse policy documents</td>
<td>Observation. Diaries. Video/audio and computer recording</td>
<td>Interviews. Questionnaires. Tests</td>
</tr>
</tbody>
</table>
The intention of this framework was to provide a variety of data and methods to be considered for particular evaluations, rather than, for example, to list the kinds of questions that are included in the checklists mentioned earlier. The overall emphasis is on educational issues and in many cases, in pursuing these issues it is necessary to observe students using the software and here it is important to try to assess how usability issues impact on learning. More detailed tools were developed at the university, including, for example an observation schedule [12] and usability measures are also used when needed. This approach is consistent with the integrative approach advocated by Squires and Preece, but is applied in the context of students’ use of software rather than teachers’ predictive evaluation.

Below we describe each of the three dimensions: context, interactions, and attitudes and outcomes, in more detail.

2.1. Context

One important aspect of context is the reason why CAL is adopted in the first place, i.e. the underlying rationale for its development and use. One long standing reason for using CAL in the OU was to provide particular support for distance learning students (e.g. to provide feedback and interaction and to reduce isolation). More recently however, there have been significant developments, e.g. home based multi-media, which may lead to different rationales for the use of new media. For example, multi-media may be used in order to bring together resources that would otherwise be separate (in video, text, audio and computer) and so give the learner an easier way of experiencing the role of primary researcher. Our evaluations have also included CAL packages designed to teach concepts in the curriculum that are known to be difficult, such as the Works Metallurgist, designed to help students with interpreting phase diagrams. Such different rationales for the use of CAL will require different evaluation approaches (addressing different questions) in order to assess the extent to which the innovation has achieved its aims. Other aspects of context include features of the setting, e.g. home or institution, the role of teacher and whether learners work individually or collaboratively.

2.2. Interactions

The reason for looking at students’ interactions with the software is in order to understand more about their learning processes. This can provide information on why and how particular elements work (or do not). That is, rather than just finding out what does and doesn’t work, it allows some analysis of that process. Where possible (e.g. at residential schools) students are observed working with the software. Such interactions can be audio or video-taped, in order to provide protocol data for later analysis. Computer logs can also be collected of all key presses and the routes that students take through the materials. There may also be products of student work (e.g. worksheets completed whilst using the package).

2.3. Outcomes

Being able to attribute learning outcomes to CAL when CAL is one part of a multi-faceted course, is very difficult, as we have discussed previously. To investigate the
educational effectiveness of a CAL program, information from a variety of sources needs to be used, including, where appropriate, pre and post-achievement tests, interviews and questionnaires with students and tutors to assess attitudes and perceptions.

3. Case studies

The CAL evaluation project conducted a number of separate but linked evaluations over a two year period: including evaluations of a tutorial on phase diagrams, a physics simulation, chemistry tutorials, art history multimedia and numeracy CAL for a foundation technology course. Three case studies are described here. Two of these were part of the CAL evaluation project, the third is a research project which used OU students as subjects but was not part of the CAL evaluation project although it used the CIAO! framework for an evaluation study. The framework was intended to be facilitative rather than prescriptive and for each evaluation, different methods were both possible and appropriate. For some, it was easier to obtain outcome data than interaction data, but the case studies were chosen to illustrate, between them, the range of methods and data, in order to assess the success of this evaluation model when applied to a variety of cases.

What follows, then, for each of the case studies, is

- a brief description of the course;
- the use and rationale for the CAL;
- the particular questions or issues under investigation;
- an account of one or more of the methods that were particularly salient in the evaluation;
- a brief summary of the findings;
- a commentary on the effectiveness of the method.

4. Case study 1: the living with technology course

4.1. The course and the CAL materials

Living with technology is a first level course, accessible to students with no previous background. By the end of the course, students should be ready to do a second level course, equivalent to second year in full time undergraduate programmes. All course blocks have an associated mainstream book along with tributary texts covering the knowledge and skills necessary to understand and analyse each issue. The residential school offers practical work including laboratory and field work, tutorials lectures and discussions. Students are responsible for making their own arrangements for access to a computer with the given specifications.

4.2. The use and rationale for the CAL

Five of the six blocks have a numeracy tributary with associated CAL software aimed to help students develop the necessary mathematical skills needed for the mainstream course.
These computer based materials are provided in parallel with text-based materials covering similar topics and students can choose the medium they prefer. All the self assessment questions are in the software and arranged as pre and post-tests. The aim here, then, was to help students to assess and develop their mathematical skills. A major part of the rationale for the pre-test approach was that there was a large population of students with very diverse backgrounds and a wide range of basic numeracy skills. The evaluation ran over two years, in the first year the CAL software ran under a DOS based graphical user interface and in the second year was rewritten to run under Windows.

4.3. The questions or issues under investigation

- How much use did students make of the Numeracy materials?
- How did CAL compare with the booklets: what were the relative advantages and disadvantages?
- What contribution did the CAL materials make to students understanding?
- How did usage compare over two years?
- What was the student feedback on new CAL packages where further changes were possible (formative evaluation)?

4.4. Salient methods

In many evaluations it is not feasible to collect a range of data across the three dimensions shown in Table 1. For our evaluation of the living with technology course, however, we were able to do this (although evaluating learning outcomes remained problematic) and a summary of the methods used and data collected is given in Table 2.

Data on student interactions was collected in the formative evaluations of the new CAL packages for Blocks 2 and 3 of the course, but only Block 2 will be reported here. The equations part of Block 2 aims to teach precedence rules in manipulating algebraic equations. The evaluation team ran an evening session on the new Block 2 equations package for local students who identified themselves as having problems with algebra.

Table 2
Framework for evaluating the living with technology course

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>Students completed tasks and were observed whilst they worked with new CAL packages</td>
<td>Questionnaires to investigate students use of CAL programs and numeracy booklets</td>
</tr>
<tr>
<td>Students subsequently filled in a questionnaire and took part in a discussion group</td>
<td>Interviews to investigate students’ attitudes and perceptions</td>
</tr>
<tr>
<td>Data</td>
<td></td>
</tr>
<tr>
<td>Notes of the problems and issues revealed during observations</td>
<td>Computer conference for students to comment on the new CAL packages</td>
</tr>
<tr>
<td>Comments from the discussion group and questionnaires</td>
<td>Students’ attitudes and perceptions</td>
</tr>
<tr>
<td></td>
<td>Students’ usage over two years</td>
</tr>
</tbody>
</table>
Students were asked to run the Equations part of the software and researchers observed and noted any problems they had and any unexpected events. At the end, students’ comments were gathered from a discussion and via a questionnaire.

With respect to outcomes usage data was collected from questionnaires over two years and students’ perceptions and attitudes were monitored via interviews at residential schools during the first year of the project when comments were also considered via a computer mediated conference (FirstClass).

4.5. Brief summary of findings: benefits and other perceptions

In a formative context observations of students’ use of the Block 2 material revealed issues and problems which would have been very difficult to find out about in any other way. For example, interface problems included a ‘drag and drop’ action to a formula box that was too sensitive and that required a precondition that was not explicit. In addition observations revealed the use of terminology that was either unfamiliar to the students or had been forgotten, e.g. ‘operator’ and a bug was also revealed.

Finally, there were discussions about how best to support procedural skills. Here students’ needs differed and they had divided views. The program works by initially directing the students through the calculations in a step by step manner then requiring students to do some part of the calculations themselves. Some students felt that they were working through the problem at a mechanical, surface level without having to consider the process and engage with it. For others it helped them to get the overall view of the problem without getting lost in the detail, for one who was very anxious, the fact that the program takes the learner through in a very gentle way without requiring too much made it possible for him to try it without his anxiety level becoming unmanageable.

From a summative perspective, usage of the CAL numeracy materials increased significantly over the two years studied. The second year of usage coincided with the move to Windows and it is likely that this move contributed to the growth in its use by reducing the interface problems.

Students’ perceptions of the CAL were positive. Students commented that they found the CAL useful for diagnosis, self testing and practice, and some students found the programs particularly helpful for revision because they could use it when there was only a limited time available. For example, one student commented that: “CAL is excellent for quick revision or when I only have a short amount of time available”.

Finally, many students used both the CAL programs and the booklets, using whichever was most convenient at the particular time they needed it.

4.6. Commentary

Having both the numeracy booklets and CAL materials gave us an opportunity to compare these media and it is interesting to note that students appreciate both, with the vast majority of the students (93%) stating that they preferred having both media. Different methods were necessary to give overall usage, comparisons, attitudes and perceptions. Most importantly, observations of software use as part of the formative evaluation enabled the designers to make changes.
One question that should be considered is whether the problems and issues discussed previously would have been revealed by predictive evaluation with teachers. It is likely that the interface problems such as the ‘drag and drop’ action would have been revealed, as would the use of unfamiliar terminology. However, the issue about the best way of helping learners to acquire the skills needed for manipulating equations is a crucial one that predictive evaluations with teachers would not have revealed, as the learners’ perceptions of this aspect of the CAL package and the extent to which they found it useful, varied and related to their confidence and previous knowledge. Moreover this is exactly the kind of issue where it would be very difficult to consider usability and learnability separately, they need to be considered in an integrated way. We might note, further, that the question of how computers can best support learners’ understanding of mathematical skills with a procedural component is part of a long standing debate, which can be rather simply stated as: does removing the tedium of the calculations from students help them focus on the problem at a more fundamental level (at a higher level of understanding) or is doing the calculations part of achieving that understanding?

5. Case study 2: A427 Charles Booth and social investigation in nineteenth century Britain

5.1. The course and the CAL materials

Charles Booth and social investigation in nineteenth century Britain is an advanced undergraduate course. It introduces students to the computer as a tool for historical research and students learn to use and construct historical databases, and to formulate their own research question making use of the data presented in the CD-ROM. Many students are expected to go on to carry out their own research at post-graduate level.

An introductory screen represents Charles Booth’s study from which all relevant categories of data can be accessed by clicking on the relevant word (e.g. biographies or glossaries). Students can access the material on the CD-ROM in different ways: they can reach Toynbee Hall, for example, from Booth’s study, from the contents, from the London Walks maps or via a related activity. Four tutorials were provided to teach students how to use various aspects of the CD-ROM.

The 46 students studying the course in 1997 were spread around the UK, making it impossible to provide regional tutorials and so all tuition for the course took place via the computer conferencing environment, except for two day schools at the Open University campus.

5.2. The use and rationale for the CAL

All the original and essential data for the course is on CD-ROM. This makes it possible to collect together essential primary data from diverse media including published writings, transcribed manuscript material, Booth’s colour coded maps of Victorian London, O/S maps, illustrations and plans, original census data, photographs, audio, biographies, bibliographies, glossaries and on-screen tutorials. One aim of the computer conferencing environment is to provide teaching that supports the CD-ROM resource.
5.3. The questions or issues under investigation

It was important to find out whether students could, indeed, cope with the technology, whether the on-screen tutorials succeeded in guiding students around the CD-ROM and the contribution of the conferencing system. The evaluation aimed to find out:

- What problems do students face in using the CD-ROM for their coursework?
- How much do they use the on-screen tutorials and how useful do they find them?
- How much was the FirstClass conferencing system used by students and what was their perception about its contribution to their studies?

5.4. Salient methods

There was a limited opportunity for face-to-face contact with students at a day school where it was possible to observe students working and also to interview four students. Outcome data was collected by a questionnaire which was sent to all students (46) and returned by 25 students. There was also a FirstClass conference for problems related to the CD-ROM as well as the on-line tutorials provided via the conferencing system which provided further interaction data. This is summarized in Table 3.

5.5. Brief summary of the findings

All students were able to install and set up the software successfully. As found in other courses using CD-ROMs [13] reading too much directly from the CD-ROM was tiring and so students printed some of the material and also requested that the instructions be in printed form.

Students did not report any serious navigation problems in accessing the different parts of the CD-ROM although some commented that it could be time consuming to go to the same place unintentionally because there were many different routes to the same material.

Table 3
Framework for evaluating the ‘Charles Booth’ course

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>Students were observed constructing historical</td>
<td>Questionnaires designed to investigate</td>
</tr>
<tr>
<td>databases at a day school</td>
<td>students’ use of the CD-ROM and also their</td>
</tr>
<tr>
<td></td>
<td>attitudes and perceptions</td>
</tr>
<tr>
<td>Four students were interviewed</td>
<td>A computer conference for problems related to</td>
</tr>
<tr>
<td></td>
<td>the CD-ROM</td>
</tr>
<tr>
<td>All computer conference messages were collected</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td></td>
</tr>
<tr>
<td>Notes of students facility with and problems in</td>
<td>Students’ use, attitudes and perceptions</td>
</tr>
<tr>
<td>constructing databases</td>
<td></td>
</tr>
<tr>
<td>Interviews</td>
<td></td>
</tr>
<tr>
<td>Messages from the tutorials which were part of</td>
<td>Students’ use of the CD-ROM</td>
</tr>
<tr>
<td>the computer conferencing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
They found working with the text-based resources easiest followed by the databases and the maps. By contrast, annotating the maps was perceived as difficult by nearly half of the students. Most students praised the CD-ROM, for example one student commented that “it brought (the) archives into immediate personal contact”. They were also positive about facilities such as the search facility that could be used to analyse Booth’s texts, the databases and the census data. However, some students found that working with the materials in the CD-ROM was time consuming because of the richness of the materials and the subsequent side tracking.

The computer conferencing included structured academic conferences for each of the two tutorial groups, along with social and ‘chat’ conferences and technical help. The feedback was very positive and nearly all the respondents used it for their studies. Whilst the majority of students used FirstClass mainly to read other messages and learn from messages that applied to their situation, nearly as many also contributed to academic conferences and two thirds used it to ask their tutor for help (these categories are not exclusive). More than half of the students believed that they had learnt a significant amount from the tutorial conferences.

Apart from helping them to feel less isolated, students also commented that the process of writing the messages helped to clarify their thoughts and some thought it would be impossible to do the course without it. The students’ contributions showed that they used the conferencing to propose and develop arguments about social investigation which both made use of the CD-ROM and which acknowledged and built on other students’ contributions. The following are two extracts from students’ contributions following a tutor’s question about women’s reasons for being involved in social investigation:

My initial response…was…yes… Social investigation did give women an opportunity to participate in the public sphere… However it then occurred to me that there’s an assumption in this question that women are an homogenous group which of course is not true. They were as separated by birth education, etc. as men were and we should distinguish between these groups. Having looked at the biographies on the CD-ROM it struck me that all the women…whilst not necessarily…wealthy, were invariably not…from the lower social classes…

I want to take up a point made by Sam (April 11th) regarding Clara Collet and ‘feminism’… I don’t think it’s possible to draw the conclusion that Clara showed ‘feminism’ tendencies just because she didn’t marry… I do agree with Sam and Sue that feminism is not an easy…term to define in this context. It is a word that is heavily loaded with C20 meaning.

Both these contributions show students engaging in a reflective debate, where the CD-ROM material is used to support, illustrate and develop academic arguments. Contributions also built on others’ contributions through challenging some aspects and supporting others as in the example earlier. Although much of our previous experience was that students appreciate the medium for keeping in touch with other students, i.e. its main function was social, this study provides support for Harasim’s [14] claims that computer conferencing supports and enables such activities.
5.6. Commentary

In this course, the aim of including the CD-ROM was to make it possible for distance learners to obtain primary data from diverse resources in order to carry out project work. It was essential that students should be able to use the CD-ROM successfully, yet most of the teaching and support to help them do so was also provided by the computer through tutorials and supported by on-line conferencing. The findings of the evaluation, summarized previously, suggest that the risky route was well justified. The combination of a small amount of observation and interviewing, plus the questionnaires allowed judgements about the usability of the CD-ROM software and revealed the students’ perceptions. Access to the conferences enabled an evaluation of both the amount of contribution made and the nature of the debate.

These methods allowed us to answer the questions posed earlier. What they do not do, however, is to allow us to address a further question which arises when a course like this using a multi-media CD-ROM and supported by computer conferencing appears to have been a great success, is this perception reflected in what students have learnt and in the quality of the students work? It would not be appropriate to administer pre and post-tests to students on a course such as this, which is a resource based course leading up to a project. A more feasible alternative is to engage in further analysis of the computer conferencing and to ask subject experts to comment on the level of academic debate and to analyse the students projects, both to assess their quality and also to see to what extent they demonstrate the use of the various resources of the CD-ROM at the students disposal. Both these activities are planned for the near future.

6. Case study 3: the Link program

6.1. Introduction

Link is part of a wider research project [15] which is investigating students’ understanding of correlation and the principled design of software to address students’ misunderstandings. It was tested out with OU students but is not currently included in any university courses.

6.2. The program

Link is a stand alone package on correlation [15], designed to be used by second or third year undergraduate psychology students. The purpose of the program is to enable students to review their understanding of correlation. The aim of the package is to ensure that students have a clear idea about the different kinds of relationships between variables.

6.3. The use and rationale for Link

Previous research has shown that students can hold a number of misconceptions in correlation, e.g. a causalistic conception and an unidirectional conception [16, 17]. The three activities in Link are designed to address particular misconceptions.
The theoretical approach underlying Link is anchored instruction [18] and so the topic correlation is presented in the context of studies in psychology. Currently, one study about TV violence and children’s aggression was implemented and in this study, students can choose to work through three activities. The program uses correlations that are similar to the ones obtained in an actual study but also provides additional negative coefficients in order to address the unidirectional misconception.

6.4. Issues under investigation

The CIAO! framework was used in the second phase of the formative study, (the first phase had investigated usability) where the aims were to: (a) pilot the tests designed to elicit students conceptions in correlation; (b) to see if the program contributes to a change in students conceptions in correlation; and (c) to formatively evaluate the program’s activities and presentation of topic material.

As the study was a formative evaluation of courseware which is not being used as part of a course, the CIAO! framework was adapted to primarily focus on two of its dimensions: interactions and outcomes and was used to organize the different kinds of data that were generated and collected in the study (Table 4).

6.5. Evaluation method

The evaluation was carried out with 18 students at a residential school who were given a number of tasks to complete whilst they worked with the program and were asked to think-aloud whilst they did this. To assess learning outcomes, two equivalent tests in correlation were developed [17] and adapted for the current study. Students completed one test before working with the Link program and another test after they had completed the tasks.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Interactions</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactions</td>
<td>Students completed tasks and were observed whilst they worked with the program</td>
<td>Questionnaires designed to elicit students</td>
</tr>
<tr>
<td></td>
<td>Students were asked to think-aloud whilst they completed the tasks</td>
<td>Equivalent tests completed by the students before and after they used the program</td>
</tr>
<tr>
<td></td>
<td>A record of observations and audio recording of students think-aloud</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Records of student interactions and think-aloud</td>
<td>Students’ opinions about the program</td>
</tr>
</tbody>
</table>

Table 4
Framework for the second phase of formative evaluation
6.6. Brief summary of findings: learning outcomes

In this case study we focus on learning outcomes (although the evaluation covered many different areas [17]. The pre and post-tests provided an overall score for each participant. A one-tailed $t$-test showed that there was a significant difference in the mean scores for the pre and post tests. Participants whole question scores provided data about their correct and incorrect answers to particular questions. Here it was revealed that there was a significant change from the pre to the post-test in the frequencies of correct and incorrect responses on questions about the strength of correlations.

Link is still under development and the activities that were implemented took a relatively short time, whilst the tests themselves took as long, so it is possible that the general improvement in test scores was caused by a practice effect. Such a possibility is not unusual in evaluating CAL. Finished programs may still only occupy an hour or so of the student’s time, and as Pea commented in reviewing Logo evaluations in the 1980s, in order for such an innovation to make a significant impact it is necessary for a relatively weak intervention to lead to a strong outcome [19].

The Link evaluation study was formative and did not include a control group, which would be needed to provide stronger evidence for the impact of the Link program and further research will address this issue.

So whilst outcome data can be collected it is still complex and difficult to ascertain what to attribute outcomes to. However, the range of methods providing outcome data, talk aloud protocols, students perceptions and interviews, has provided very helpful information at this stage.

7. Review of model

We believe that the CIAO! model has enabled us to focus on what is most salient for particular applications. Over all evaluations, two areas have proved difficult, learning outcomes and observations.

There are well known difficulties in documenting learning outcomes that include issues such as the desirability and possibility of comparative studies and establishing that any differences in outcome are particularly attributable to the CAL (for example Draper et al. [20] and Jones et al. [9]). Part of our reaction to this problem lies in the CIAO! model, which attempts to provide information about the whole teaching and learning context by focusing on different parts of the process and products. It will only be appropriate and possible to collect information about learning outcomes in particular cases. For example, in the living with technology case study, the students completed a computer marked assignment early on which assesses their numeracy ability. Some areas of the course exam require that they apply their numeracy skills effectively. This could provide a naturally occurring correlation experiment where we could look at the students whose scores were low on their initial assignment and see whether those who used the CAL scored better on their exams than those who did not. However, eliciting learning outcomes data requires fine judgement both in terms of whether such a question (e.g. has the student
learnt X better/more quickly as a result of using CAL package Y?) is the most appropriate to ask in the particular context and whether and how it can be answered.

A closely related issue is whether it is appropriate to compare the use of CAL to some other media. Where CAL is replicating other media or face to face teaching, comparisons may be possible and appropriate (this was the case with the living with technology numeracy materials, which were produced in two forms) but where new technology is used to teach or provide some experience that cannot otherwise be provided, it is not clear what the comparative base should be and the question of whether it is ‘more effective’ or ‘teaches better or faster’ is not the appropriate question to ask.

Educational software is changing. Instructional and diagnostic software will always be valuable but recent developments have focused much more on using new technologies in resource based learning, which Taylor and Laurillard [21] define as “open access, self-directed learning from a large information source”. Often, such software does not have any assessment. The rationale in some cases might be to replace other forms of teaching, or it may be to provide access to various resources. This alone will not make it a learning resource, several issues need to be considered, including assumptions about students’ skills in learning with resource based materials, how students navigate the material and the difference between ‘instructional’ and ‘information dissemination’ systems [22].

According to Romiskowski [22] in an instructional system the teaching component presents information, evaluates the student’s performance and feeds that back into the interaction to improve the quality of learning. By contrast, information dissemination systems do not assess whether learning has taken place but focus on content, organization, structure and presentation. Both the usability and the educational effects of these complex environments are crucial.

The ultimate objective for educational software is that it should be educationally beneficial and it is exactly in such environments that it is important to track how usability contributes (or not) to educational goals. In their argument for an integrative approach, such as using the Jigsaw model, Squires and Preece [1] support a distributed view of cognition and maintain that it is not possible to evaluate an educational software application predictively without reference to a perceived educational setting. Computer assisted learning, like any learning, takes place in a setting which includes framing the task, so it is hard for a teacher who is not involved in that setting to make predictive evaluations that are meaningful. Some learning and usability issues will be discovered by teachers, others will not. We argue that the next stage of involving the learners themselves is crucial. There is much that is unknown about the use of multimedia, as well as recognized problems (such as navigation) so it will be crucial to track students use of resource based packages very closely to uncover the problems and successes and although observing and interviewing are time consuming methods, they will need to play a large part.

Acknowledgements

We would like to thank our many colleagues who contributed to the various case studies reported here: Rosemary O’Day, David Englander, Tony Anthoni, Peter Cox, Karen
References

[12] Open University, Evaluation Methods and Procedures for studying learners’ use of media, developed by PLUM (Open University) and TELL (University of Hull) see http://www-iet.open.ac.uk/iet/PLUM/Contents.html, 1996.