Monitoring distance education students' practical programming activities

Dr Pete Thomas
Department of Computing
Open University
Milton Keynes, MK7 6AA
United Kingdom
p.g.thomas@open.ac.uk

Carina Paine
Department of Psychology
Open University
Milton Keynes, MK7 6AA
United Kingdom
c.b.paine@open.ac.uk

ABSTRACT
Students studying a distance education course on introductory object-oriented programming are provided with a set of practical activities to support their learning. Each practical activity consists of a set of tasks to be worked through at home over a specified period of time. Whilst students are provided with considerable guidance on how to approach their practical work we have no knowledge of their actual behaviour. Therefore, we set up a project named AESOP to monitor what students actually do. This paper describes the AESOP system for recording and analysing the work done when undertaking the practical activities. Results have been obtained from an analysis of data from 30 students studying the 1999 presentation and 200 students on the 2000 presentation about their attempts at 8 sets of practical work. We found significant differences between the course designer’s perception of student work and how students actually behave. Students usually needed significantly less time to complete their practical work than expected indicating that their perception of the difficulty of the tasks set is often different from that of the course designers. We also found that there is a tendency for students to limit their interaction with the computer to about 30 to 40 minutes.

Keywords
Distance learning, Programming, Student study behaviour, Electronic observation

Introduction
The Open University and the AESOP educational environment is one in which 5000 students study independently at a distance, off line, using software developed for an entry level, distance education course in computing. The course M206, Computing: An Object-Oriented Approach (M206, 2000) involves practical work using Smalltalk in the LearningWorks environment (Goldberg et al., 1997). Students receive a CD containing the LearningWorks system and a collection of LearningBooks. A LearningBook, which we henceforth abbreviate to LB, consists of a set of software modules of objects and classes based on a book metaphor, and form the units of practical work. Students submit a small number of assignments spread over the year long duration of the course. No one knows how a student has interacted with a LB that is not explicitly assessed in one of the assignments. Interest is growing in the area of understanding precisely what students do (Biggs, 1999) and people are beginning to question whether the model of a compliant learner (one who complies with the instructions and expectations of the teacher) holds for the modern student (Goodyear, 1999).

AESOP (An Electronic Student Observatory Project) (Thomas et al, 1998; AESOP, 2001) is an on-going research programme into how students learn to program. A development of the project has been an Electronic Observatory, a system for recording, replaying and analysing user actions in the LearningWorks environment. The project’s main aims are to (i) investigate how students learn to program (in Smalltalk) (ii) develop software tools to help investigate how students learn to program (iii) develop software tools to help the teaching and learning process.

The notion of recording students’ actions is not new (Smith et al., 1993; Kivi et al., 1998), but in our environment in which students work at home, we wanted a mechanism that would produce small files that could be easily and cheaply downloaded from students’ machines thereby avoiding large files that would result if we were to record keystrokes. The Electronic Observatory consists of four major software components: a recorder, a
replayer, an analyser and a Coach. As a student interacts with a LB the recorder writes a file of textual representations of events e.g. button clicks, hyperlink selections and expression evaluations. Each event is time-stamped. The replayer plays back a recording so that an investigator can observe what the student did. It has been implemented as a LB application that sends instructions to an executing LearningWorks system (Thomas et al., 1998, Macgregor et al., 1999). The analyser enables the investigator to search for patterns of behaviour across sets of recordings. A wide variety of analysis tools have been identified and several built (Thomas and Paine, 2000). The Coach is a software component, also implemented as a LB, which runs concurrently with the recorder and other LBs. The aim of the Coach is to provide additional help to students studying programming, particularly in the area of error reporting and error diagnostics (Thomas, 2001).

In this paper we focus on the AESOP analyser and in particular the analyser’s Tasks Completed Tool which examines the amount of practical activity a student carries out in a LB. The motivation for this research was to discover the extent to which students complete the practical work set for them. While the design of the course is such that practical activities form a significant component, anecdotal evidence from other computing courses suggests that students do not complete all practical activities set. This is particularly the case when the available time for study comes under pressure, that is, students abandon practical work in favour of studying the core written materials. Therefore, we wished to investigate the extent to which students engage with the practical work and whether there was any correlation between work done and performance on the course. Such information would be beneficial to both the course designers, in maintaining the course, and students, by providing evidence of the value of doing the practical work.

The Tasks Completed Tool

The recordings of students’ work, whilst small in terms of file size, contain a great deal of information which makes them unsuitable for extensive human evaluation. As we were aiming for a large-scale trial that would involve several thousand recordings, an automatic means of analysing them was required. The Tasks Completed Tool analyses recordings of students’ practical activities and compares the tasks that students have completed successfully (tasks completed) with the tasks they were asked to complete (tasks set). A task is a small grained programming activity, such as entering a fragment of Smalltalk code for the LearningWorks system to evaluate. Figure 1 illustrates that a practical consists of a number of tasks, and that collections of practicals are designated as sessions. For pedagogical reasons, students are expected to complete a session in one sitting, that is, using the computer without a break. A LB is a collection of one or more sessions. A discussion includes solutions and explanations of the tasks set in a practical. Students are encouraged to attempt a practical before turning to the associated discussion.

Figure 1: The structure of a LearningBook

To specify the tasks contained in a LB, we adopted the notion of an ideal recording – a complete attempt at the activities in a LB – that would be used as a benchmark against which student attempts would be measured. This information was used to create a specification of the tasks to be completed in each practical. The Tasks Completed Tool performs a comparison between the specification and each student recording. The output gives the number of tasks a student was asked to complete in a practical or discussion, and the number of tasks completed successfully. Figure 2 shows an example of the output from the tool when analysing LB 09, session 1, practical 2.
Chapter 09 session 1 practical 2
Number of set tasks = 5
Set tasks:
gribbit up, 2 colour, gibbit right, gribbit hone, htonnoy right

Number of tasks completed = 4
Tasks completed:
gribbit up, 2 colour, gribbit hone, htonnoy right

Figure 2: Example output from the Tasks Completed Tool

The tasks set and the tasks completed are listed in the output for ease of comparison. The tasks can be completed by the student in any order, but the tool will only recognise tasks completed in the practical specified.

Some tasks can be identified as completed without dispute in situations where the comparison between what the student did (recorded as a string) and the task specification (another string) is an exact match. Many tasks cannot be identified as easily. These tasks include evaluations of Smalltalk expressions and the construction of Smalltalk methods, where there may be any number of legitimate differences between a student recording and the specification. As a first step, we use wildcards in the specification, which allow for some of these differences, including differences in white space, argument values, and the naming of instance variables. For example, if a student were asked to create an instance of the class Frog, the specification would be, *:= Frog new, making allowance for the student to use any variable to refer to their Frog instance.

In some practical activities, students are asked to investigate or experiment by devising their own solutions to problems. Again, the use of wildcards in the specification can identify some of this extra work. For example, if students are asked to experiment with the colour : method the specification would be *colour:*. In effect, the specification simply demands that the string colour: should appear somewhere within the student attempt. Clearly, this scheme does not allow for all possibilities, and the tool will generally underestimate the number of tasks actually completed. However, the nature of the practical activities, particularly at the start of the course, is often one in which the expected responses are heavily constrained and the number of possibilities is quite limited.

Research questions

We were interested in whether or not students completed all the tasks set and, if not, was there any pattern in their behaviour. We had noticed, from an initial reading of some recordings that students did not seem to complete all the tasks set and, in LBs with multiple sessions, tended to try the first session and then either give up or attempt all sessions in the LB.

Table 1 shows the percentage of tasks completed for a sample of students in LB 09 (the first LB to contain any meaningful programming tasks). LB 09 is split into 4 sessions with a total of 21 practicals and 83 tasks. Clearly, the percentage of tasks completed in each session decreases as the student progresses through the LB. This data was gathered during the 1999 presentation of the course.

<table>
<thead>
<tr>
<th>% Tasks Completed (LearningBook 09)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1 Practicals 1 - 6</td>
</tr>
<tr>
<td>75.00%</td>
</tr>
<tr>
<td>Session 2 Practical 7 - 9</td>
</tr>
<tr>
<td>65.56%</td>
</tr>
<tr>
<td>Session 3 Practicals 10 - 15</td>
</tr>
<tr>
<td>62.70%</td>
</tr>
<tr>
<td>Session 4 Practicals 16 - 21</td>
</tr>
<tr>
<td>54.90%</td>
</tr>
<tr>
<td>Complete LB Practicals 1 - 21</td>
</tr>
<tr>
<td>66.28%</td>
</tr>
</tbody>
</table>

Table 1: Percentage of tasks completed in LB 09

The behaviour of students on other LBs is not so dramatic, but some patterns have emerged. Table 2 shows the average percentage of tasks completed for all LBs analysed (we looked at only those LBs that asked students to perform some programming activities). Typically each LB is expected to be studied over four hours, but for a few LBs less time is allocated.
Table 2: Summary data of tasks completed in each LB

The variable number of student responses in Table 2 simply reflects the number of recordings received from student volunteers. Some of the information in Table 2 is shown graphically in Figure 3.

Two features of Table 2 and Figure 3 are immediately apparent. First, on average, students complete over 60% of the tasks set in all LBs. Second, while some students (under 10%) complete all the tasks set, in general the majority of students do not do everything they are asked to do, and we were interested in trying to find out why.

Figure 4 shows a diagram of the number of tasks completed plotted against the number of tasks set. The upper line indicates the maximum number of tasks that could be completed (the number of tasks set). The regression line has a slope of 0.796, which suggests that, across all LBs, students complete around 80% of tasks set.

Figure 4: Number of tasks completed compared with number of tasks set
There are two natural questions to ask:

1. Does the absolute number of tasks set in a LB affect the amount of tasks a student completes in a LB? That is, is the setting of more tasks off-putting to students? Perhaps the apparent volume of work puts students off. We hypothesise that the more tasks there are to do in a LB, the proportion of tasks completed in the LB will be lower.

2. Is the time spent studying a LB related to the number of tasks set in the LB? It would seem obvious that the length of time spent studying would be related to the amount of work set, but we wondered whether the perceived amount of study required affected the students’ willingness to complete all the activities set. We hypothesise that the time taken to study a LB is proportional to the number of tasks completed in the LB.

Results

Starting with Question 1, we looked at the number of tasks set and the percentage of tasks completed in each of 8 LBs. Figure 5 shows a diagram of the percentage of tasks completed plotted against the number of tasks set in each LB.

![Figure 5: Percentage of tasks completed compared with number of tasks set.](image)

Spearman’s Rho correlation test applied to the data of Figure 5 yields a value for $r_s$ of -0.0476, when the critical value for $r_s$, for a one-tailed test when $N = 8$, is 0.643 (95% confidence). This shows that there is no correlation between the proportion of tasks completed and the number of tasks set.

To answer Question 2 we performed a Spearman’s Rho correlation test to compare the number of tasks completed in a LB with the time spent in a LB. This yielded a value of $r_s = -0.42857$, when the critical value for $r_s$ for a two-tailed test when $N = 8$ is 0.643 (90% confidence). Once more, no significant correlation was found. However, looking at the proportion of tasks completed compared with the time spent in a LB reveals a different picture, shown in Figure 6.

![Figure 6: Percentage of tasks completed compared with time spent in each LB](image)
This time Spearman's Rho correlation test shows a strong negative correlation ($r_s = -0.9048$), when the critical value for $r_s$ for a two tailed test when $N=8$ is 0.881 (99% confidence). This suggests that the more time a student spends doing the tasks, the smaller the proportion of tasks set are completed.

A possibly more revealing factor is that a LB is normally split into several sessions and earlier work (Thomas and Paine, 2000) showed that students tend to complete a session in one sitting, that is, they normally work through a session continuously. Therefore, looking at sessions in a LB might provide a finer grained view of student behaviour.

Returning to Question 1 we looked at the percentage of tasks completed in each session of a LB to see whether there was a relationship between the number of tasks set and the proportion of tasks completed. As was true at the LB level, there was no significant relationship. From these results we reject the hypothesis that the more tasks there are to do in a LB the proportion of tasks completed in the LB will be lower.

Turning once again to our Question 2, a plot of the number of tasks set against the time spent studying a session is shown in Figure 7.

For the data shown in Figure 7, Spearman's Rho test produces a value of $r_s = 0.7708$, when the critical value for $r_s$ for a two tailed test when $N=22$ is 0.562 (99% confidence), confirming a significant relationship. The regression line has a slope of 0.2349. The slope of the regression line excluding the LBs that have one session (LB12 and LB 20) is 0.4571. This illustrates that there is a stronger relationship between LBs with multiple sessions, which suggests that student behaviour when there is only one session in a LB is different from their behaviour when there are multiple sessions.

### Task difficulty

The initial analyses assume that all tasks are of the same complexity either in terms of the time taken to complete or degree of difficulty. However tasks can vary in levels of difficulty, and this is a potentially important point when comparing different LBs. We devised three metrics for task difficulty as follows:

A. the average time (minutes) spent in a LB per task set;
B. the average time (minutes) spent in a LB per task completed;
C. the recommended time (minutes) to be spent in a LB per task set.

Metrics (A) and (B) are based on student behaviour whereas metric (C) is based upon the course team's perception of the tasks. Values for the three metrics applied to 8 LBs are shown in Table 3.
As noted earlier, the tasks completed tool will have usually underestimated the number of tasks actually completed, which means that metric A provides a minimum value and metric B a maximum value. Together, metrics A and B provide a range within which the ratio of tasks attempted in the time spent lies. The data in Table 3 is also shown in Figure 8 which graphically illustrates that the course team generally over-estimates task difficulty (the middle column in each category). It is interesting to note that the two LBs in which the course team’s estimate is lower than the actual time taken by students are LB 12 and LB 20, which are the only two LBs comprising a single session.

<table>
<thead>
<tr>
<th>LearningBook</th>
<th>Metric A</th>
<th>Metric B</th>
<th>Metric C</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1.75</td>
<td>2.53</td>
<td>2.89</td>
</tr>
<tr>
<td>10</td>
<td>1.84</td>
<td>1.93</td>
<td>5.58</td>
</tr>
<tr>
<td>12</td>
<td>1.51</td>
<td>1.91</td>
<td>1.51</td>
</tr>
<tr>
<td>13</td>
<td>1.13</td>
<td>1.25</td>
<td>3.64</td>
</tr>
<tr>
<td>14</td>
<td>1.33</td>
<td>1.48</td>
<td>3.33</td>
</tr>
<tr>
<td>15</td>
<td>0.86</td>
<td>0.93</td>
<td>2.07</td>
</tr>
<tr>
<td>16</td>
<td>3.03</td>
<td>3.70</td>
<td>5.45</td>
</tr>
<tr>
<td>20</td>
<td>3.19</td>
<td>4.96</td>
<td>3.85</td>
</tr>
</tbody>
</table>

*Table 3: Values of three metrics for task difficulty*

The time spent on an individual task averaged over individual learning books shows that, approximately, students spend between 1 and 4 minutes per task whereas the course team felt that they should be spending between 2 and 5 minutes. The difficulty of each session is shown in Figure 9.

More revealing is a comparison between the task difficulty and the percentage of tasks not completed in a LB shown in Figure 10.
Spearman’s *Rho correlation test* yields a value for $r_s$ of 0.762, when the critical value for $r_s$ for a two-tailed test when N=8 is 0.738 (95% confidence), confirming that there is a correlation between the number of tasks not completed and the task difficulty.

### Time students are prepared to devote to practicals

Each session of a LB was designed to be completed in a single continuous interaction with the computer. A question naturally arises as to the extent to which students followed this pattern given the variation in the number of tasks per session.

If we define a *sitting* as a period of time that a student interacts continuously with the computer without a significant break in activity, we can determine whether the number of sittings varies with the number of sessions in each LB. The identification of sittings within each recording is not straightforward because a recording is simply a list of time-stamped events. We cannot be certain whether the time elapsed between two events represents a break in study, or an activity, such as reading, which is related to the practical activity being undertaken. Therefore, we analysed the recordings in such a way that any gap between two successive events greater than a specified threshold was taken as a break in study and divides one sitting from the next. For example, a gap greater than one hour would be indicative of a break in study. Figure 11 compares the results of this analysis for three different gap sizes: 30 minutes, 15 minutes and 10 minutes. For a gap size of 10 minutes or more, the number of sittings exceeds the number of sessions. A gap size of 15 minutes or more usually gives an average number of sittings greater than the number of sessions. A gap greater than 30 minutes results in an average number of sittings that is normally smaller than the number of sessions. The two notable exceptions to this pattern are LB 12 and LB 20 where it is very clear that students do not complete the LB in a single session as recommended, but take 3 or 4 sittings.

![Figure 11: The number of sittings per LB](image)
In earlier work (Thomas & Paine, 2000), we took any gap of 10 minutes or more to be an indication that a student has broken off from their studies. This was deduced on the following basis. Since the material required to perform individual tasks is contained wholly within the LBs, there is little incentive to interact with other sources of material, and the amount to be studied to perform a task is quite small and only likely to take a few minutes. Our conclusion is that, on average, for LBs with multiple sessions, students both follow the instructions precisely and study each session in a single sitting, or take one or two more sittings.

Figure 12 plots the time per sitting with the same three indications of study breaks. The indication here is that there is a limit to the amount of time students are prepared to spend in a single sitting. For the 10-minute gap, the limit is around 25 minutes, for the 15 minute gap, the limit is around 30 minutes. LB 20 is the exception, but even here the limit is around 40 minutes.

It seems reasonable, therefore, to suggest that sessions should be studied in a single sitting designed to take not more than 30 minutes.

The investigations discussed so far were repeated in a much larger experiment based on the 2000 presentation of the course in which we looked at the recordings from 200 students. This provided almost 2000 recordings to analyse but spread over a larger number of LBs. The data confirmed that a time difference between successive recorded events of 10 minutes yields a consistent view of what constitutes a break in study. Table 4 shows the range and average study time for a LB based on the 2000 data.

The amount of time that students spend working with each LB is substantially different from the time recommended by the course team. For the 2000 data, the average was 9%, compared with 58% for the 1999 data). The exception was LB12 where students spend significantly more time than recommended (140% in 1999 and 190% in 2000). LB12 has been identified by the course team as one that, “falls short of achieving its aims and learning objectives” (Rapanotti, L. & Griffiths, R., 2001).

The 2000 data also confirmed the view that most students attempt more than just the first session in a multi-session LB, and that the number attempting later sessions decreases towards the end of the LB. Table 5 shows the data for LBs 06 to 16 (similar results hold for LBs 17 to 39). The exception is LB15 which shows an increase.
in activity in the final session compared to the first, but this is almost certainly due to the fact that the final session is associated with an assignment.

<table>
<thead>
<tr>
<th>LB</th>
<th>No of Sessions in LB</th>
<th>No of students in first session</th>
<th>No of students in last session</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5</td>
<td>124</td>
<td>122</td>
<td>98.38</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>142</td>
<td>105</td>
<td>73.94</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>125</td>
<td>119</td>
<td>95.20</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>133</td>
<td>127</td>
<td>95.49</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>122</td>
<td>117</td>
<td>95.90</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>117</td>
<td>110</td>
<td>94.02</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>108</td>
<td>109</td>
<td>100.93</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>106</td>
<td>102</td>
<td>96.23</td>
</tr>
</tbody>
</table>

*Table 5: Students completing first and last sessions in LBs*

**Student performance**

The assessment of the course is in two parts: the assignments taken regularly throughout the course and a final examination. The course lasts for 30 weeks and the data reported upon here were collected from activities carried out during the first third of the course. Thus, the interaction with LBs forms but a part of the course as a whole with only two out of 9 assignments and a small proportion of the final exam assessing this work. Comparison of the performance on the relevant assignments and the final exam with the number of tasks completed failed to show any significant correlations.

However, the nature of the assignments is such that students are asked to carry out specific tasks and generally students will endeavour to complete all such tasks to gain maximum marks so we would not expect to find a correlation between assessed performance and overall accomplishment in tasks completed. We are currently looking for ways of distinguishing between tasks completed for an assignment and those completed in the normal course of study.

The final examination covers the whole course and again we would not expect to find a significant correlation between exam performance and the number of tasks completed. However, we shall be looking at performance on specific questions in a later study.

**Future work**

The version of the tasks completed tool reported upon in this paper has a number of weaknesses. In its initial implementation, the tool looked for an exact match between the specification and the student recording, but has been extended to allow for some variations in student responses. Nevertheless, the tool does not identify all attempts at tackling a task. Where possible we have made allowance for this deficiency by looking at tasks set. Nevertheless we wish to improve the tool’s accuracy.

A further limitation of the initial implementation of the tool is that it analyses student activity whilst they were working with the practical activities. There is evidence to suggest that some students perform the tasks having previously looked at the discussions where solutions to the tasks are given. Thus, in some cases, the evidence of tasks being completed is located in a different part of the recordings, though we believe this to be a minor effect.

In some practical activities students are asked to complete the same task more than once, but the tool does not deal satisfactorily with this situation.

Clearly, the major weakness of the present tool is that it does not directly identify whether any unsuccessful attempts were made at a task. However, the task difficulty metrics indicate that the tool provides a good estimate of the tasks attempted which gives us reasonable confidence in the conclusions we are drawing. Nevertheless, we wish to improve the accuracy of the tool and work is underway to devise and implement better ways of specifying tasks.

An area that we wish to investigate in more detail is how students approach their assignments. The LBs we have investigated to date have asked the students to engage with a series of well structured but small-scale tasks. In
assignments, students have more freedom of choice and it will be interesting to see how the analyses discussed in this paper compare with student behaviour in a less constrained environment.

Conclusion

The results obtained have shown that students generally make an attempt at a majority of activities set, but seldom complete them all. The more time they spend on the early parts of a practical activity, the more likely they are to stop before completing a significant proportion of the work set. For course designers the implication is to avoid overly long sequences of practical activities otherwise there is a significant likelihood that students will not engage with the later material.

There are significant difference between the course designer’s perception of student work and how students actually behave. Students usually spend significantly less time on their practical work than expected by the course designers and their perception of the difficulty of the tasks set is often different from that of the course designers. However, within any group of students, there will be a range of times taken to complete the work and our course designers take the view that over-estimating the average time required enables a large proportion of the student body to tackle the practical work within the allocated time. Of greater concern is where insufficient time has been allocated for the average student. The results reported here enable the course designers to give more accurate advice to students.

Perhaps the most significant result is that students are prepared to spend up to about 30 to 40 minutes at a time tackling practical work. We have traditionally tried to design our distance education courses in ‘chunks’ such that each chunk could be studied in a single evening (for between 2 and 3 hours). The results reported here seem to indicate that students are not prepared to spend the whole of this time sitting at a computer. However, it is useful to note that, having decided to tackle practical work, students can be relied upon to spend up to half an hour doing the work. In other words, it is probably best to provide at least half an hour’s worth of practical work per session and avoid smaller amounts (one can conjecture that breaking off from normal study to do the practical work must be seen as worthwhile otherwise students will view it as unimportant).

These findings have informed the course designers and are being taken into account for the 2002 presentation. In particular, the teaching in LBs 09 and 16 has been revised. More significantly, our results that relate to the time students are prepared to engage with practical activities are being used by the developers of a new level one course in computing.

Acknowledgements

Our thanks to Malcolm Macgregor for his diligence in implementing the Recorder and the Analysis tools and insightful comments on the research in general.

References


