Evaluating OASYS

Ashley Ward (ashley@dcs.warwick.ac.uk)

October 2000

Copies: Warwick Teaching Certificate * 2
AHB, SBR

Summary

The reasoning behind a questionnaire designed for the evaluation of a Computer Science module is given. Some simple analysis of results from questionnaire responses is presented, along with some deeper analysis involving correlation with other available data. The results are used as a basis for suggested recommendations for change.

Introduction

Design of Information Structures (CS126) is a core module for first year Computer Scientists, Computer Systems Engineers, Computer and Business Studies students and Computer and Management Science students, and is optionally available to students on other degree programmes. It aims to introduce learners to some of the fundamental "building blocks" in Computer Science – the abstract data types of lists, stacks, queues and so on. The programming language used for the more practical aspects of the module is Java, so the module has the secondary aim of helping learner's advance in their expertise with this language (which a large proportion of learners will have studied for one term only so far). The module is currently taught using a combination of lectures, practical coursework assignments, an examination and laboratory sessions.

Each student on the module should attend four lab sessions. In each session, they spend 90 minutes following exercises presented via an on-line worksheet, and then take a test related to material that they have just encountered. Each test runs under exam conditions for 30 minutes. The 1999 entry encountered the tests in a web-based form, driven by a system we named OASYS (for Online Assessment SYStem). The system presents learners with both multiple-choice questions and open questions to which free answers are given. A novel feature of the system is that free answers given by learners are assessed by other learners in a form of automated peer assessment. OASYS has been the subject of other presentations [Ward2000a], [Ward2000b], [Ward2000c], [Bhalerao2000] – this essay will concentrate on evaluation of the system, primarily through analysis of a post-module questionnaire. Some preliminary analysis was recently presented at ALT-C 2000. My slides and notes from that presentation (which may provide a helpful overview of the system and process) are included in Appendix (xii).

This essay is intended to fulfil dual roles: it should both provide useful evaluative feedback to staff involved with CS126 and OASYS, and also demonstrate my achievement of the relevant learning objectives for the Assessment and Evaluation module within the Warwick Teaching Certificate.

Methodology

After a full term's use of the system, I began to consider a formal evaluation of our work, and designed a questionnaire. I hoped to add the information gathered by the questionnaire to the large amount of data already present in the OASYS database of learner's responses, where both their answers to tests and their peer assessment of others are recorded. Compared to analysing the data represented in a large stack of paper scripts, obtaining statistics from the data within OASYS is relatively easy (if a little more technically involved!).

Due to the wealth of situation-specific data available, I decided to broaden the questionnaire a little in an attempt to discover some more general results. The final questionnaire therefore can be divided into four different sections, asking questions which attempt to discover:

1. the identity of the respondent in the hope that the questionnaire response can be linked to the data already available,
2. the respondent's approach to study in general (questions 1-18),
3. the individual's evaluation of their experience from the lab sessions in responses to closed questions (questions 19-35),
4. any other issues that the individual wishes to raise in response to more open questions (questions 36 and 37).

Section 2 is drawn from a questionnaire included in [Gibbs88], which in turn is a simplified version of a questionnaire taken from [Entwistle88]. The questions included can be split into categories, which attempt to score the respondent on the following scales:

A. 'achieving orientation'. "The extent to which students are competitive, well organised and concerned to do well".
B. 'reproducing orientation'. "The extent to which students are attempting to memorise the subject matter".
C. 'meaning orientation'. "The extent to which students are attempting to make sense of the subject matter".

National norms to which populations can be compared are given in [Entwistle88]. I also intended to correlate the results with the other, context specific, questions asked and other data from the database.

The resulting paper form questionnaire is included in Appendix (i). I decided to use a paper-based form, to be completed within a lecture rather than an on-line variant as we were now in the last week of the second term, and I thought it unlikely that many students would spend time logging on to assist at this stage.

Analysis

The results of my analysis are included in Appendix (ii-xi). The work has drawn on two years of exam results, coursework marks, lab test data and questionnaire responses. The Appendix shows analysis in several sections:

- A statistical summary of the measured population's approach to study (results from section 2, shown in Appendix ii).
- Statistical summaries of the population's evaluation of the lab sessions (results from section 3, shown in Appendix iii).
- Categorised listings of the free responses given in section 4. (Appendix iv, v).
- Graphs and statistics of lab, coursework and examination result distributions from 1998/1999 and 1999/2000. (Appendix vi, vii, viii).
- Correlation figures between all available variables matched up on an individual basis, presented both as scatter plots (Appendix ix), raw results (Appendix x) and explained in English (Appendix xi).

The Appendix is intended to be self-explanatory. However, I will now place some of the more interesting results in context.

Comparing our questionnaire respondent's A B and D scores to Entwistle's norms shows our population to be moderately keen (they have a high A score). The questionnaire was distributed and answered during the last CS126 lecture of the term, and only 59 responses were received, compared to a figure of over 200 students participating in the module. We could surmise that only the most motivated students turned up to this last lecture. The scores may also have been affected by a design error in my layout of the form: the fact that the central '2' response should only be used in exceptional circumstances seems to have escaped the attention of many respondents.

During the second term of 1999/2000, we ran 12 DIS lab sessions, and each session had four demonstrators in attendance. Leaving aside the test, each session provided a possible one and a half hours of contact time between

demonstrators and learners. Following this and assuming the demonstrators spend all of their time in contact with the learners, each learner should receive on average 20 minutes with a demonstrator. Q20 ("…I spent this amount of time receiving help from a demonstrator") gives an average of 9 minutes, so our students are not receiving quite as much contact as they might.

The average responses to Q21 ("I did the on-line worksheets outside the labs") and Q23 ("it was clear that I had to mark 3 scripts…") are disappointing: less people did the on-line worksheets outside the labs in 1999/2000 than in the previous year, and the amount of marking that they were required to do was not clear to around a third of the population. The average response to Q25 ("… I spent this amount of time marking") shows that even the keen population that responded to the questionnaire only spent around 5 minutes marking each script. Given that some tests contain around ten free-answer questions to be marked and that useful commentary feedback takes a little time to compose, this figure seems a little low. Results from Q29 ("I received speedy feedback on my work in the tests") are also disappointing, if expected – 56% of the population did not feel that they received speedy feedback on their work in the tests. This is due to the concurrent building and running of the test system: the feedback interface was not actually implemented until quite late on, and so some results took several weeks to arrive.

Some of the questionnaire design did not work out as expected. For example, on reflection I now consider Q22 ("if the labs had not been assessed, I wouldn't have done them") to be misleading and hence invalid. The wording of Q22 was copied from a 1998/1999 questionnaire in order to compare the results across years – but what exactly does a negative answer mean? Q33 ("getting full credit for the lab sessions is important to me") and Q34 ("it would be worthwhile cheating in the tests if possible") in particular may look a little strange at first glance, but they are intended to measure whether the respondent perceives the lab tests as summative or formative. The tests were envisaged as primarily formative, but a (very small) summative mark is given to them in order to encourage active attendance at the lab sessions. This mixed decision, or perhaps

an overly-subtle questionnaire design, shows up in the results, which are inconclusive as to the students' perception of the process as summative or formative.

More encouraging news can be found in the responses to Q28, which go some way to justifying the use of peer assessment in the system: 90% of respondents realised mistakes they had made in their own answers whilst marking. Also of some interest are the responses to Q32, which show that 94% of respondents feel that anonymous marking of the tests is important or are indifferent to the question. This can be compared to anecdotal evidence from the TELRI project [TELRI], which also involves a variant of peer assessment, where Mick Roach has informally mentioned the *lack* of need for anonymity during a seminar. This might be due to the fact that TELRI case studies have so far mainly concentrated on the social sciences, where learners rarely pick the same essay topic, and hence answers are not as directly comparable as in OASYS.

The free response answers raise two topics strongly in the questionnaire. Students felt that the lectures and lab sessions should be more related, and also requested more time. The content of the lab sessions was the same in 1999/2000 as in the previous year, when this point was not raised – perhaps the change was in the lectures, where possibly the lab sessions were less explicitly referred to the second time around.

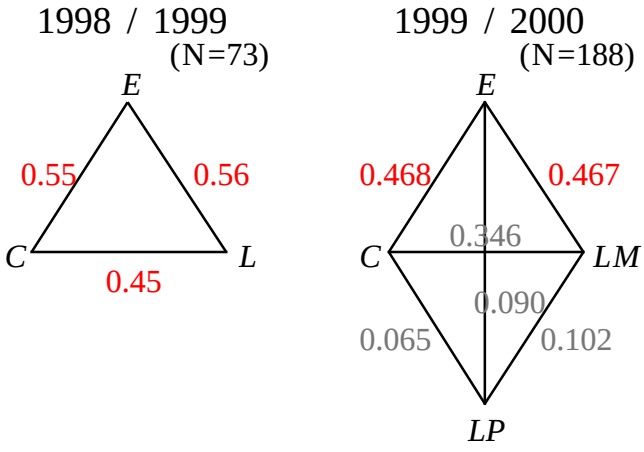
Comparing the lab mark distributions from the two years for which data is available is an interesting exercise. The 1998/1999 distribution is presumably more valid as the lab scripts were then marked by "expert" demonstrators and staff. The 1999/2000 marks were generated by peer assessment (although marks were moderated by staff if multiple peer marking showed disagreement or if moderation was requested by a learner). The two distributions are similar in shape and have a very similar average value (when comparing the 1998/1999 results with the 1999/2000 results including participation). However, no marks above 76% were given during peer assessment, leading to a much less spread distribution. The relatively harsh peer marks given may have led to increased

stress among the population and possibly to the requests for more time mentioned in the free responses.

Generating correlations between all the possibilities mainly gives quantitative evidence backing already existing intuitively felt qualitative conclusions. Interesting results amongst the mass of expected results include [lab mark / Q24 0.59], which appears to show that the more marking a learner does, a higher peer-assessed mark they can expect to receive themselves. Also [lab mark / B –0.43] shows that learners with a strong reproducing orientation tended to receive lower marks, which could be considered a positive result if this module is intending to encourage deeper understanding of the material. Learners that tend to attempt to memorise material also stated that they found marking difficult, as evidenced by [Q26 / B –0.45].

The correlation links between exam results, coursework marks and lab session grades is shown graphically in the figure below. In 1998/1999, links existed between all three results. Lab grades for an individual were a reasonable predictor for examination result, and also, but less significantly, for coursework marks. In 1999/2000, where the lab marks (LM) were produced mainly by peer assessment, they were still a significant predictor of examination result, but not this time of coursework mark. Despite their practical emphasis, the lab tests may still be testing skills more useful in an examination than in practical assignment building and documentation.

E = exam
C = coursework
L = labs
LM = lab mark
LP = lab participation



Conclusions

In this section I draw on the results to present some possible recommendations for the system.

Faster feedback – encourage them to mark more effectively

- The feedback interface is now in place.
- Better communication with the students – perhaps a handout describing the process, their expected input and return and perhaps some carefully selected results from the data presented here.
- Could they be encouraged to mark in pairs?
- Learners could be publicly ranked in marking effectiveness terms, in an attempt to harness their competitive nature, in a similar way to other websites (such as the Search for Extraterrestrial Intelligence at home [SETI]). Perhaps a prize could be awarded for the "best" marking.
- Participants could be automatically mailed reminders about marking.

Link lectures and labs

- More reference to the labs within lectures.
- Perhaps the material needs more co-ordination: but then again, this issue did not arise in 1998/1999, so perhaps the linkage simply needs to be more explicitly stated.

Summative or formative?

We need to ensure active attendance, but even a small amount of module credit appears to result in the students behaving as if this was summative assessment.

- Perhaps full credit could be given once a minimum standard is reached, with merit awards for very high achievers to encourage work beyond the minimum.

Improve quality of marking

- Can we mark the marking? This could be done by asking learners to give feedback on their received marks. Alternatively (or additionally), perhaps

each set of markers' marks could be normalised (assuming they mark a good sample of the population) and the amount of correction required could be used to give the system feedback on the quality of their marking.

- "Trust" algorithms which trace paths from initial "seeds" in the system (expert markers in this case) are possible.

Simplify the overly complex marking collation strategy

Due to the many possible cases of unanswered, unmarked, not-enough-marked, fully marked, moderated, marked-on-one-criteria-only (etc) states for each answer, the process of collating marks together to form a result for a script is overly complex and may well contain a bug. It is also difficult to tell when the fully correctly marked state is reached and how far off this is.

- Peer assess individual questions and ignore the "script" (a set of one learner's answers to one test) concept during the marking process.
- Don't attempt to collate a summative mark for an entire script, simply leave it as a collection of individual responses to individual answers.

Better system reliability

- Perhaps run the system on its own machine, with minimal dependencies on other infrastructure. It would still need to be regularly backed up.
- Simplify the access privileges – could they do the tests in their own time, whenever and wherever they like?
- Run the tests all in one building to aid communication between students, demonstrators and system administrators.

Reduce stress / more time please

- Simplify the (previously skewed and complex) timetable.
- Show personal learner statistics in context of overall population results.
- Integrate the "test" into the lab session.

Many of these suggested changes would take some significant time to complete, which simply is not available at the moment. However, the very existence of the system seems an improvement over a paper-based system and the value of peer assessment has been partly validated in the answers to Q28 where 90% of the population reconsidered their answers on marking others. It is hoped that the analysis presented here will be useful and lead to improvements in the application of OASYS.

References

[Bhalerao2000] Abhir Bhalerao and Ashley Ward (2000) *Towards Electronically Assisted Peer Assessment: A Case Study* Paper in preparation for submission to the forthcoming special issue of ALT-J.

[Entwistle88] Noel Entwistle (1988) *Styles of learning and teaching* David Fulton Publishers, London.

[Gibbs88] Gibbs, Habeshaw and Habeshaw (1988) *53 Interesting Ways to Appraise Your Teaching* Technical and Educational Services Ltd.

[SETI] The Search for Extraterrestrial Intelligence at home `http://setiathome.berkeley.edu/`

[TELRI] Technology Enhanced Learning in Research-led Institutions `http://www.telri.ac.uk/`

[Ward2000a] Ashley Ward (2000) *A Case Study in Electronically Assisted Peer Assessment* Essay submitted towards completion of the module Preparing to Teach for the Warwick Teaching Certificate.

[Ward2000b] Ashley Ward (2000) *Toward electronically assisted peer assessment: a case study* Presentation P027 given at ALT-C 2000 on Monday 11th September 2000.

[Ward2000c] Ashley Ward (2000) *OASYS profile* A small website containing the above essays and other information: `http://www.dcs.warwick.ac.uk/~ashley/Research/OASYS/`

Appendix

i. DIS laboratory sessions development questionnaire
ii. Summary of responses to questions 1-18
iii. Summary of responses to questions 19-35
iv. Summary of responses to question 36
v. Summary of responses to question 37
vi. 1999/2000 DIS lab result distributions
vii. 1999/2000 DIS exam and coursework result distributions
viii. 1998/1999 DIS exam, lab and coursework result distributions
ix. Scatter graphs of 1999/2000 data used for correlation
x. Raw correlation results
xi. Correlations explained
xii. "Toward electronically-assisted peer assessment: a case study" – slides and notes from a presentation given at ALT-C 2000.