Does it Make a Difference if Students Exercise on the Web or in the Classroom?

Ari Korhonen and Lauri Malmi
Department of Computer Science and Engineering
Helsinki University of Technology
Finland
{archie,lma}@cs.hut.fi

Pertti Myllyselkä and Patrik Scheinin
Department of Education
University of Helsinki
Finland
{Patrik.Scheinin,Pertti.Myllyselka}@Helsinki.FI

ABSTRACT

Several Web-based learning environments which can automatically give immediate feedback to the students have been reported within the past few years. The quality of feedback can be relatively high in these systems, but it does not achieve the level a trained teacher can provide. However, the lack of the best possible feedback can be compensated for, to some extent, by providing immediate and individualised feedback at any place or time. The question is whether the systems providing automatic feedback are good enough to compete with humans.

This paper reports on a randomised large scale intervention study. We found that there was no significant difference in the final examination results between students doing instructed simulation exercises in a classroom session and students using a web-based learning environment, if the exercises were the same. However, with more challenging exercises, there was a significant difference in the examination results, while the drop out rate was higher. Thus, the chosen teaching method and medium did not effect the level of learning, but the quality of the exercises did.

Categories and Subject Descriptors

K.3.1 [Computer Uses in Education]: Computer-assisted instruction (CAI), Distance learning; K.3.2 [Computer and Information Science Education]: Computer science education—*data structures and algorithms*

General Terms

Experimentation, Human Factors

Keywords

Automatic assessment, Randomised intervention study

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ITiCSE'02, June 24-26, 2002, Aarhus, Denmark.

Copyright 2002 ACM 1-58113-499-1/02/0006 ...\$5.00.

1. INTRODUCTION

Several attempts have been made to produce both effective and high quality environments that are capable of automatic assessment of exercises [1, 2, 3, 4, 5]. The need for automation is due to the trend towards teaching large numbers of students at low cost. Moreover, virtual courses benefit from automatic feedback since it is difficult and expensive to provide round the clock service with human resources.

Recent automatic assessment systems are not capable of assessing students' exercises in as much detail and giving as sophisticated feedback as teachers do. On the other hand, some of these systems have advantages that could compensate for these limitations. In our study, the underlying system [4] has features that traditional instruction cannot provide. The students solve *individually tailored* exercises by means of algorithm simulation¹, and the system is capable of *immediate* automatic assessment of these exercises *around the clock*. The assessment produces feedback that enables *revision of answers*. Moreover, after the deadline has passed, the system automatically produces *individual model solutions* for the exercises.

There is a need for research that estimates the quality, advantages, and limitations of these approaches. The question is whether the advantages of fully automatic assessment can compensate for its limitations? In this experimental intervention study with three randomised groups A, B, and C ($N_A = 372$, $N_B = 77$, $N_C = 101$), student performance was monitored over the second year course in data structures and algorithms during a twelve week period. The examination results of students in the virtual learning environment were compared with those in the traditional classroom sessions.

The results show that, if the exercises are the same, there is no significant difference in the final examination results between students exercising on the Web and in a classroom. However, if the exercises are more challenging, there is a significant difference in the examination results, but additionally the drop out rate is higher.

We describe the research design used in this survey in Section 2 in detail. Section 3 summarises the main results obtained during the research. Finally, in Section 4 we present the conclusions.

2. RESEARCH DESIGN

The course consists of three separate and compulsory parts which are homework exercises, project, and examination,

¹This method is described in more detail in Section 2.

Gr.	N	s	ave	Teaching method
A	372	1	372	virtual simulation exercises
B	77	3	25	classroom simulation exercises
C	101	2	50	classroom design assignments
Σ	550	6	-	-

Table 1: Summary of Groups. N denotes the total number of students in a group. Each group was split into s subgroups of average size ave . Each group had a different teaching method; this was the variable of the survey.

which have weighting of 30%, 30%, and 40% of the final grade of the course, respectively. The homework exercises, described below in more detail, included small algorithm simulation, design, and analysis exercises which students had to submit every two weeks during the course. The project included a larger, fairly complicated design assignment which was solved in groups of 2-3 students at the end of the course. The project assignment was not a target of this research, and we do not discuss it here more closely.

For this research, the students were divided into three separate groups, A , B , and C , each of which had different homework exercises. The experimental group A used the web-based learning environment with automatic feedback and grading of (virtual) simulation exercises which has been used at Helsinki University of Technology since 1991 [4]. The control groups B and C had traditional classroom exercise sessions in which the tutors gave feedback for the students about the exercises. The control group B did the same simulation exercises as the experimental group A . However, the control group C did exercises which were beyond the scope of automatic assessment (*i.e.* designing new algorithms, writing essays, *etc.*). This latter set of exercises was designed to be more challenging for the students than the simulation exercises of the web-based learning environment.

The groups were formed randomly. The sizes of the control groups were limited by the classroom capacities. Group B was divided into 3 subgroups with approximately 25 students in each, and group C into 2 subgroups with approximately 50 students (Table 1). Each subgroup had two tutors who gave feedback to the students during the classroom sessions. Group A was not split since they worked with the web-based exercises and space was therefore not an issue. All students had the option of getting guidance from course assistants during office hours, or by asking questions in the course newsgroup.

All students in this study were freshmen on the course, *i.e.*, students who had failed the course previously as well as foreign students were not included in the study. Such students were, however, allowed to take the course.

2.1 Exercises and assignments

The homework exercises included algorithm simulations exercises solved in the web-based learning environment, algorithm simulation exercises solved in classroom sessions, and design assignments solved in classroom sessions.

2.1.1 Simulation exercises

The simulation exercises were the same as those we have been using since 1991 in our Data Structures and Algorithms course. The automatic assessment system used in these exercises is capable of producing an individual input data structure for each student. The student is asked to simulate an algorithm with a given input and to report one or more subsequent states of the data structure. For example, the input data structure might be a set of keys that the student is asked to insert one at a time into an initially empty AVL tree. The set of keys are chosen in such a way that both single and double rotations occurs. Here, we want to test whether the student has understood how the AVL tree insert operation works. As an answer, the student produces the predefined state(s) of the target structure. The correctness of the AVL tree can be assessed, for example, after 6 insertions (*i.e.* at least one single rotation occurs), and again when all the keys have been inserted (*i.e.* at least one double rotation occurs).

The automatic assessment system is capable of producing a model solution for each student by running the actual implementation of the algorithm. The model solution is then compared with the given answer, and the feedback reveals how many steps are correct. The grading phase uses several heuristics to convert this feedback into points. For example, some minor errors might be allowed without the loss of points. Certain exercises additionally provide several correct model solutions. For example, there are two implementations for a Binary Search Tree insert operation: one that inserts duplicates to the left and one that inserts them to the right.

Groups A and B completed the same set of simulation exercises. Group A received the exercises and sent the answers back to the system either by email or by using a Web-based graphical front end application which is provided for solving the exercises in a graphical form. The system sent immediate feedback to the students by email. Based on this feedback, a few iterations were allowed in which it was possible to improve one's performance by revising incorrect solutions. Usually, the students had two weeks to deal with one set of exercises. There were six sets of exercises. After the deadline of a set, the system sent the individualised model solutions to the students by email.

In group B , the exercises were the same but the tutors gave the feedback at the classroom sessions. The exercises for group B were not individually tailored; all group members had the same input for each exercise. Revision of one's solution was allowed within the two-hour classroom sessions allocated for group B . Of course, it was suggested that the students should do most of the work at home before they came to the classroom. In the classroom sessions, some of the students did the rest of the exercises in pairs but the exercises were individually assessed. Each session dealt with one set of exercises. Thus, there were six sessions for each subgroup every other week.

2.1.2 Design assignments

The homework exercises for group *C* were more challenging than those described above. We did not only cover the algorithms described in the text book, but also designed new algorithms solving similar problems to those in the simulation exercises. For example, if the text book described the standard Tree Traversing algorithms for such binary trees, where each node has links to the left and the right subtrees, the assignment was to design the same set of algorithms for the common tree implementation, in which each node has links to its first child and its next sibling.

The working method in the classroom was similar to that of group *B*, but the subgroups were twice as large. Those students who did at least some of the exercises at home before the classroom session usually benefited most from the feedback. They additionally formed study groups of two or three more often than students who did all the work in the classroom sessions. Usually, this latter group of students did not receive full points from any set of assignments.

2.1.3 Examination

The final examination was designed to cover all subject matter. There were five kinds of assignments for each main topic (sorting, searching, graphs) of the course. First, certain key concepts had to be defined. Second, an important data structure or algorithm within the topic had to be defined or explained. The students had a degree of freedom in deciding which structure/algorithm they explained, *e.g.*, they could define any balanced search tree instead of a certain balanced search tree. In the third assignment, they had to apply their algorithm to a simulation case, where the initial data structure was given. Fourth, they had to explain how algorithms within the field can be compared with each other while choosing an algorithm for an application. Finally, they had to construct some simple new algorithms and analyse their performance briefly.

3. RESULTS

The correlation between the exercise points and the examination points is high in each group (Pearson's $r_A = 0.402$, $r_B = 0.420$, and $r_C = 0.577$). Thus, there is significant positive linear relationship between the two variables. More close examination reveals that there is also significant differences between the groups.

The results of the study are summarised in Table 2. The three groups were compared with each other in respect to two important factors. The performance of a student was determined according to the *examination points* of the course. Not all of the students who initially enrolled in the course were able to finish the course successfully at the final examination. Consequently, the *degree of drop out* during the course was additionally measured. The drop out in this case, however, was determined by counting those students who were not able to finish their exercises before the end of the course. On the other hand, not all of the students that were able to finish their exercises showed up in the final examination.

Note. The students had in total 5 possibilities to take the course examination after the course. The examination points are reported only from the first examination, in which most of the students attended. The points from the latter examinations are not reported since they are not comparable with each other. On the other hand, the drop out rates

Gr.	<i>E</i>	<i>D</i> ₁	<i>D</i> ₂	<i>D</i> ₃	<i>P</i> _{exam}
<i>A</i>	33.04	4.6%	8.9%	4.8%	82%
<i>B</i>	32.74	0.0%	12%	5.2%	82%
<i>C</i>	39.56	8.8%	26%	2.0%	64%

Table 2: Summary of Results. *E* denotes the average examination points (max = 70). *D*₁ and *D*₂ denote the percentage of students who dropped the course in the beginning without any submissions, and those who dropped the course during the exercises, respectively. *D*₃ denotes the percentage of students who passed the exercises but either did not attend any examination or failed the examinations. *P*_{exam} denotes the percentage of students who passed both the exercises and the examination.

are reported after all the examinations because the comparability of the examination results is not important from this point of view.

3.1 Performance in Final Examination

Those students of group *C* who attended the final examination did well compared with the other groups. The t-test showed that their performance was significantly better than that of groups *A* and *B* (*p*-values 0.000 and 0.005 respectively). On the other hand, there is no significant difference between groups *A* and *B* (*p* = 0.871). However, the comparison becomes complicated if the huge drop out in group *C* is taken into account. Examination of the data shows that there is no single attribute that could account fully for the phenomenon.

3.2 Drop out

There are several reasons for the higher percentage of students who failed the exercises in group *C*. First, the groups knew each other; thus, it was natural that they compared their exercises. The design assignments were more difficult than the simulation exercises, and some students in group *C* felt that we pushed them too hard, which caused lower motivation. Second, the design assignments may have been too difficult, causing the students to feel anxious and worried. Third, the size of the group was possibly too large for only two tutors to provide enough feedback for such a difficult set of problems to as many as 50 students within the two hours time limit. Moreover, those who took part in the examination appeared to perform better in all of their studies than those who did not.

Within groups *A* and *B* the number (*D*₁ + *D*₂) of students who dropped the course during the exercise phase is about the same. On the other hand, the drop out in the very beginning of the course is different. Students attending the classroom exercises in group *B* were motivated to start the course seriously although a number of them gave up later. Students working in the web-based environment dropped the course easier already in the beginning, whereas those who started working seriously were less likely to give up later. This phenomenon was much stronger within the group of students (not presented in the table) who were not taken into the research setting. 45% of these students, among which were those who had failed the course earlier, foreign students and a number of working adult students, dropped the course either before or during the exercises phase.

4. CONCLUSION

First of all, there is no significant difference in the learning results of students doing their homework exercises in a virtual learning environment and students attending a traditional classroom. In addition, the commitment to the course (low drop out), is almost equal in both versions of the course. However, students working in the virtual environment seems to be more likely to drop the course earlier than students working in the classroom.

The examination results were much better when the exercises were more difficult. The problem is, however, how to maintain a reasonable drop out rate. The results imply that, if we intend to improve learning, we may choose either traditional or web based learning assignments, as long as they are suited to the subject at hand. We should, however, pay more attention to the quality of the exercises. Furthermore, the quality of the feedback and the time used for dealing with a topic should be appropriate for the challenges to the students. If the exercises are more challenging, more support and time are required to cope with the task.

It seems that the challenges in our virtual learning environment are not difficult enough. At least the improvements we have made over the past few years, (*e.g.* the possibility to revise answers, graphical front end *etc.*) have made it possible for students to cope with the exercises much better than, for example, five years ago. Thus, the future plan is to produce a more powerful system that is capable of delivering more challenging exercises, if and when the student has shown himself/herself capable of handling the basic exercises.

5. REFERENCES

- [1] S. Benford, E. Burke, E. Foxley, N. Gutteridge, and A. M. Zin. Ceilidh: A course administration and marking system. In *Proceedings of the International Conference of Computer Based Learning*, Vienna, Austria, 1993.
- [2] J. English and P. Siviter. Experience with an automatically assessed course. In *Proceedings of The 5th Annual SIGCSE/SIGCUE Conference on Innovation and Technology in Computer Science Education, ITiCSE'00*, pages 168–171, Helsinki, Finland, 2000. ACM.
- [3] D. Jackson and M. Usher. Grading student programs using assyst. In *Proceedings of 28th ACM SIGCSE Tech. Symposium on Computer Science Education*, pages 335–339, San Jose, California, USA, 1997. ACM.
- [4] A. Korhonen and L. Malmi. Algorithm simulation with automatic assessment. In *Proceedings of The 5th Annual SIGCSE/SIGCUE Conference on Innovation and Technology in Computer Science Education*, pages 160–163, Helsinki, Finland, 2000. ACM.
- [5] R. Saikkonen, L. Malmi, and A. Korhonen. Fully automatic assessment of programming exercises. In *Proceedings of The 6th Annual SIGCSE/SIGCUE Conference on Innovation and Technology in Computer Science Education, ITiCSE'01*, pages 133–136, Canterbury, United Kingdom, 2001. ACM.