Evaluation of Distance-learning Environments: Impact of Usability on Student Performance

Tselios N., Avouris N., Dimitracopoulou A., Daskalaki S. (University of Patras, GR)

(International Journal of Educational Telecommunications, 7(4), pp. 355-378 (2001)

Experimental results of usability evaluation of a distance learning system are presented in this paper. An experiment is described that took place in the frame of a University course. The main goal of the experiment was to evaluate the usability of the Testing and Self-evaluation component of the system. A complementary research goal was to explore the eventual impact of system usability on student performance. For this purpose, two alternative software components were compared that shared similar functionality, implemented in different ways (IDLE, WebCT). The usability evaluation was based on user questionnaires. From this experiment correlation between the software usability and student performance has emerged, underlining the importance of usability evaluation of systems supporting distance learning.

Recent years have witnessed the development of new powerful enabling technologies related to distance and collaborative learning. Advances in networks performance and the widespread use of the Internet made it possible for educational material of high quality to

become available to large numbers of potential learners. Additionally, these technological advances have accelerated the development of educational material for distance learning, offered through the web. Most Universities and other educational institutions engage the web in their traditional everyday activities and offer educational material of various forms for distance learning to a wider extramural audience. Yet this new use of computer technology in the educational field raised once more skepticism on the effectiveness of the process (Fitzelle & Trochim, 1996).

The world-wide web (WWW) is the technological environment that enabled and supported this process. There are many reasons for which the web can be considered a suitable educational medium: It is easily accessible by many groups of learners. It supports multiple representations of educational material and various ways of storing and structuring this information. It is powerful and easy to use as a publishing medium. Additionally, it has been widely accepted that the hyper-medial structure of the web can support learning. Some researchers characterize the web as an active learning environment that supports creativity (Becker & Dwyer, 1994). According to (Thuring, Mannemann, & Haake, 1995) the web encourages exploration of knowledge and browsing, behaviors that are strongly related to learning. The associative organization of information in the web is similar to that of human memory and the process of information retrieval from the web presents similarities to human cognitive activities. However a hyper-medial space, like the web, cannot be considered, only by these features, as an effective tutoring environment. It is rather more appropriate to think of the web as a powerful tool that can support learning, if used in an appropriate way (Eklund, 1995; Alexander, 1995). This is because learning is a process that depends on other features, like learner's motivation, previous experience and learning strategies that the individual has been supported to develop, etc.. Effectiveness of any educational environment cannot be considered independently of these aspects. It is widely accepted that effective learning is also related to educational environments and tools that provide the students with incentives for

active participation in the learning process. So the characteristics of the tools used to support learning are factors affecting the process. One of the most important features of any software tool is its *usability*, that is the effectiveness, efficiency and satisfaction that gives to its user in a given context of use and task. So the usability of an educational environment is related to its pedagogical value (Kirkpatrick, 1994) and evaluation of its usability is part of the processes of establishing its quality. However, evaluation of usability of a distance-learning environment is not an easy task. The effectiveness of usability evaluation techniques varies, depending in great extend on the specific characteristics of the evaluated environment and the objectives of the evaluation study (Molich et al., 1999). Some of the most widely used techniques are heuristic evaluation (Nielsen, 1993; Levi, Conrad & Frederick, 1998), field studies and observation (Togniazzini, 1992), questionnaires filling, interviews, logging of user performance in laboratory conditions, etc.

While there is a large corpus of theoretical and practical knowledge relating to software usability evaluation in general and educational software in particular, see (Squires & Preece, 1999; Avouris, Tselios & Tatakis, 2001), there are not established techniques relating to distance-learning environments usability evaluation (Heines, 2000). This is due partly to the fact that distance learning is an area of relatively short history, characterized by rapidly shifting technological context and by inherent idiosyncrasies of the environments under evaluation. For instance, users of distance-learning tools, in contrary to traditional software, can access them through various computer and social contexts, the process of logging their performance and actions presents technical difficulties, the rate of novice users is relatively high, while in general the characteristics of typical users of distance-learning services cannot be easily predicted. According to (Hayes, 2000) usability evaluation of online course delivery systems should examine in particular the effort required by the user to take ownership of the system's functionality and should concentrate on ease of use. It should be mentioned here, that other areas of web-based applications and tasks like information and multimedia content

distribution and e-commerce applications seem to have similar problems as far as usability evaluation is concerned, according to (Nielsen, 2000).

An Overview of the Paper

The research reported here is part of the effort to delineate and expose some of these problems through a specific case study involving usability evaluation of a module of a distance-learning environment, used under realistic educational conditions. A distance learning software environment contains usually a number of components with different functionalities. Modules that are used for content presentation, student communication with tutors and peers, collaboration and interaction support modules, modules for active learning etc. One of the components that are encountered most often in these environments is the *Testing and self-assessment* component. Such a module is usually simple in terms of functionality and design of interaction. It contains a number of closed questions with pre-determined set of answers. User interaction and user tasks are trivial and therefore one should expect that usability in this context is not an important issue. So usability assessment of such components is not normally performed due to the conventional and predictable character of the tools involved.

In the frame of our research, concerning evaluation of distance-learning systems, one of the objectives has been to establish a methodology that includes suitable techniques for evaluation of the various components of distance-learning educational environments and relates effectiveness of the tools to their usability. This approach has involved an extensive evaluation experiment of a distance-learning environment in use in the Department of Electrical and Computer Engineering (ECE) of the University of Patras, developed by our

group, the *Infotronic Distance Learning Environment (IDLE)*[*] (Avouris & Tselios,1999). This is an environment developed over the last years and actually in operation, supporting students of the ECE Department and the general public in studying a number of computer and electronics-related subjects. This paper presents results related with a usability evaluation experiment of the *Self-assessment and Student-testing* component of the system.

Additionally, the effect of system usability on student performance was studied during this experiment. This was made possible by measuring the performance of the students using the system during three different sessions. The fact that this part of the study was concentrated on a specific component of the system has made it possible to clarify methodological aspects of the usability evaluation process and relate the usability parameters studied to system functionality. One of the most interesting conclusions was that even in the case of this simple module, system usability affected student performance. The paper presents the methodology used for evaluating the system; the results of the evaluation experiment and includes discussion of the effect of usability on student performance. The experiment involved a number of students who used the module under realistic educational conditions. Some measures had to be taken during this experiment in order to control the uncertainty of distance-learning conditions: The students were collocated in the same laboratory and used similar equipment and network bandwidth. The task was transformed from that of self-evaluation that is usually performed with this module, to a testing task in order to make sure that identical conditions of use were applicable to all students and the performance of the students involved was measured. An absolute and a relative measure of usability were established. The latter involved evaluation of two alternative systems with similar functionality. Interesting results emerged from this comparative study, relating to the impact of system usability on student performance.

[*] Infotronic Distance Learning Environment has been built in the frame of the European Research Project CBT-Kernel part of the Leonardo Programme of the EU, with the participation of Fraunhofer-Institut Integrierte Schaltungen, Universität Erlangen-Nürnberg, University of Patras, MicroLEx Systems A/S

The presented results are interesting to researchers and practitioners involved with development and evaluation of distance learning technology and to the growing number of educators who are concerned with the educational effectiveness of distance-learning services and educational material provided to their students.

CONTEXT OF THE STUDY

This study concerned evaluation of the *Student Testing and Self-evaluation Component* of the Infotronic Distance Learning Environment (IDLE). IDLE (www.ee.upatras.gr /cbtkernel) has been developed over the last years in the ECE Department of the University of Patras in order to support educational activities primarily of the Department students.

During the Academic year 1999-2000 the distance-learning component went in operation and a number of curriculum subjects, mostly in the area of computer science were included. IDLE has been used experimentally to support the students of the courses 22Y103 (Introduction to Computers) and 22C901 (Data and Knowledge Base Systems) of the ECE Department, while material is currently being developed for more subjects, including Microelectronics and VLSI design.

Basic components of the system are the hyper-medial content presentation component, the component of student peer interaction, the common bulletin board and the student testing and self-evaluation component. Also support for the tutors is provided, as tools are included for tutors to develop and integrate new material and monitor students' performance. The tutors can also link content to self-evaluation material and establish alternative interaction flows for the students.

IDLE has been developed using Active Server Pages (ASP) technology that permits linking of underlying databases to dynamic web pages for user access to the content. Templates exist at the server side to which specific content, like a quiz question is loaded at

run time according to the interaction requirements. The system is browser-independent and no special client-side software is required. Users of IDLE are monitored during their interaction with the environment. The educational material visited by the student, the testing and self-evaluation questions answered, the time spent in components of the environment are stored in the user model. The users can inspect at any time the information stored about themselves, in particular their performance and history. The IDLE environment also comprises tools for student access to bulletin boards, threaded messages tool and a message broadcasting facility. The tutor can inspect at run time the student community, hold a discussion with them, look into their performance records etc.

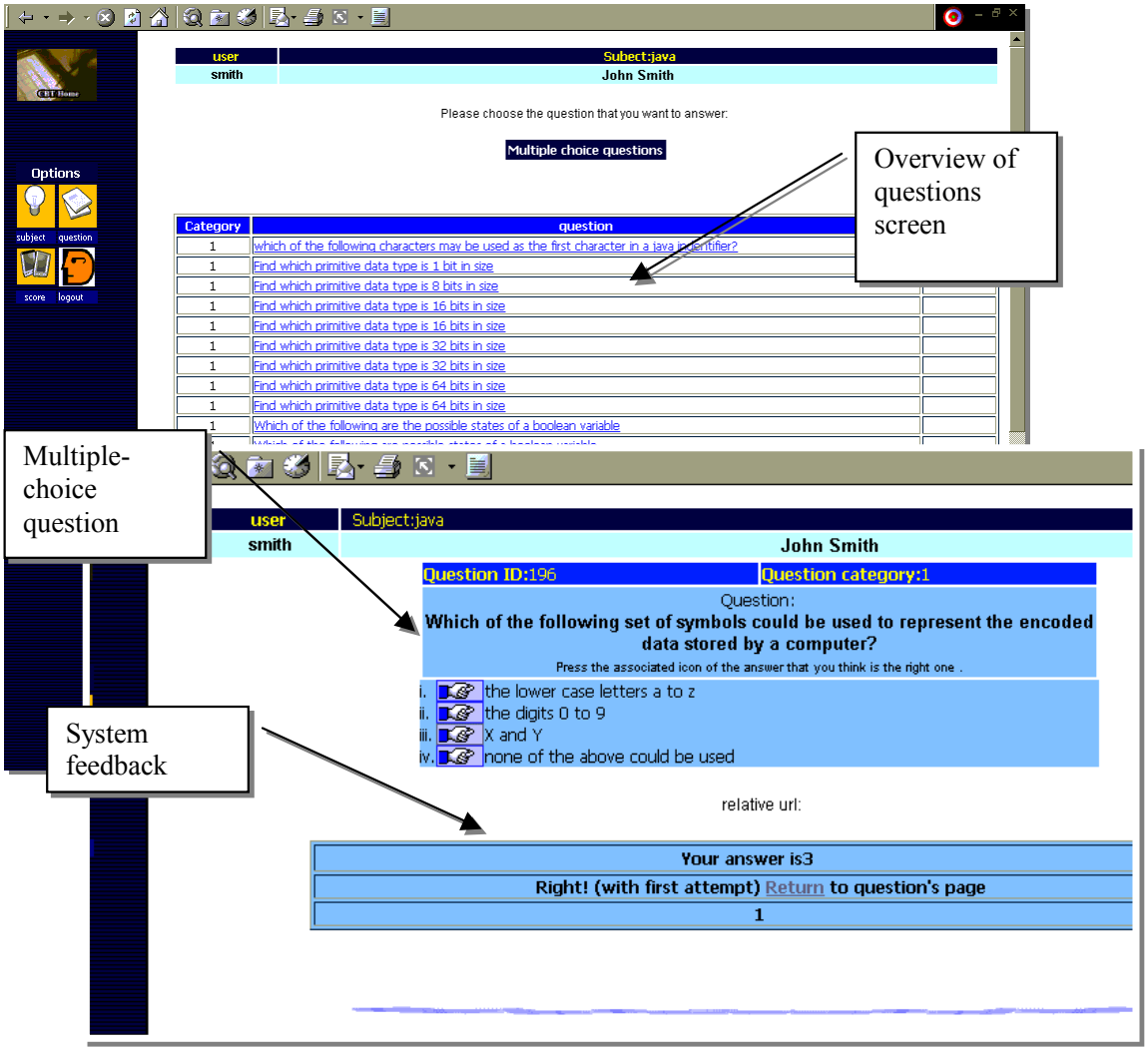


Figure 1. IDLE Testing and Evaluation Module: The questions overview page and a typical multiple-choice question page.

The module of IDLE that has been the subject of evaluation during the reported research is the *Testing and Self-evaluation Module*. This is a special area of the system related to specific subject matters contained in the system. The student can select the module usually as a means for self-assessment of his/her progress. The introductory page of the module contains an overview of all the contained quiz questions, represented as a list of short descriptive phrases. The student can select one question from the list and enter a specific page where the question is presented to the user, as shown in figure 1. Two specific kinds of quiz questions are supported by the system: multiple-choice and fill-the-blank questions. Feedback is provided to the student according to the selected answer. Multiple attempts are allowed, according to the number of the available choices.

The interaction of a student during a typical session with this self-assessment module involves move from the question overview page to specific question, work at the single question level, where the question can be answered by selecting one of the proposed answers or filling the blank, depending on the type of question. Occasionally the student can move to the score overview page in order to examine the progress of the self-test. In this page, information about the number of questions answered and the overall score (percentage of correct answers) is provided in a graphical way. In the overview page, information about the visited questions is provided by the color of the relevant hyper-links. However the student has no indication on whether a certain question has been answered or just visited, since as it is known, the semantics of hyper-links cannot be defined at this level of detail. Also a usability problem of IDLE relates to the delay observed every time a student visits a question page. The content has to be loaded to the client, so some delay can be observed depending on the network performance. So navigation though the testing material is not fluent as in a paper and pencil environment in which the student can glance through the entire exam paper before concentrating in certain questions.

The evaluation involved comparative evaluation of the IDLE tool to an alternative distance-learning environment, used as reference. This second environment was WebCT©, (Goldberg, Salari & Swoboda, 1996), a product of similar functionality to IDLE, widely used for authoring and delivering distance learning courses. A brief presentation of this product is attempted here. WebCT (Web Course Tools) has been developed initially by the University of British Columbia for supporting web-based learning. It became a product in 1997 and many modifications and enhancements have been produced since. Version 2.1 has been used in our experiment. WebCT is based on CGI (Common Gateway Interface) technology. Perl, Javascript και Java are used for creation of a virtual classroom. The educational material can be structured in sequencial and hierarchical form, while indexing and glossary facilities are provided. The educational material can be presentation slides, HTML pages, text documents

and other media. Support for asynchronous communication (bulletin board and e-mail) and synchronous (chat and virtual meetings) is also included. Finally, quizes and tests can be prepared containing questions of various kinds, while processing of the students answers and presentation of the results in various forms to the students concerned and the tutors are also supported .

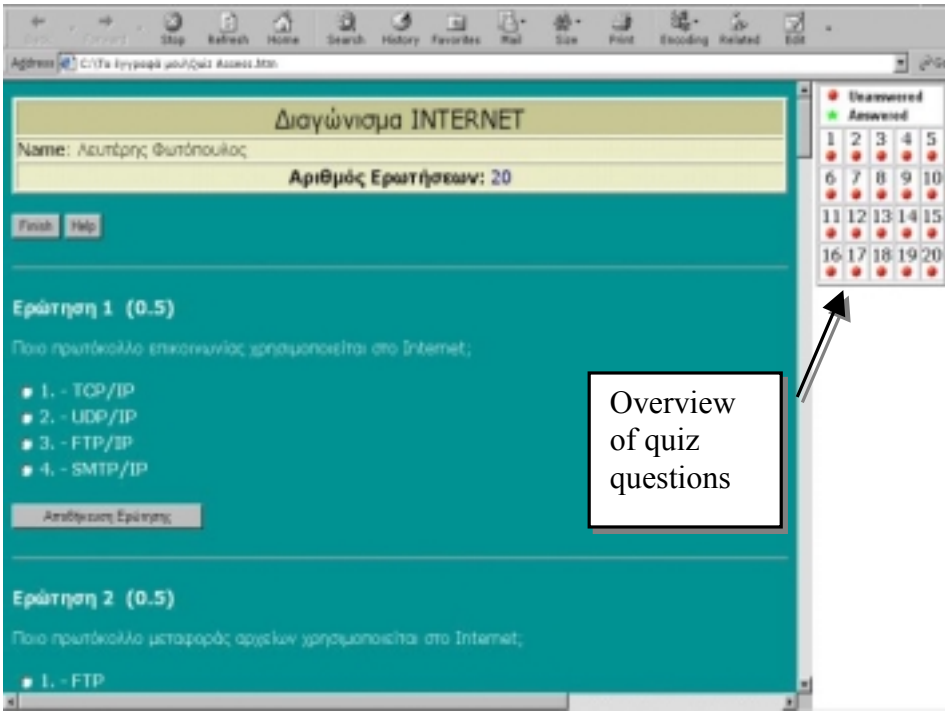


Figure 2. Typical quiz interface for the WebCT environment.

In order to perform the comparative evaluation of IDLE and WebCT, described in the following section, we developed, using WebCT, a distance learning site containing the same educational material with IDLE. However while the educational content of the two environments was identical, student interaction with them presented considerable diversity, due to the different design of the user interface of the two alternative software environments. The quiz in the WebCT environment was organized as a long page containing all the multiple-choice and the fill-the-blank questions. So the students could obtain quickly an overview of the entire exam. The students who used this environment had to scroll down the quiz page in order to navigate through the questions. So, in contrary to the IDLE solution, there was no

delay due to download time of every quiz question. An overview of the progress was provided by a table in which the answered questions are marked in a different way to the unanswered ones, as shown in figure 2.

The typical interaction with the quiz material involved in this case concentration in a question, selection and submission of an answer and moving down to the next one using the scroll-down handler. If the subsequent question was not one that the student could easily answer, scrolling further down in the exam was the immediate reaction. However, in contrary to IDLE, no feedback on the correctness of the selected answer was provided. Another difference between the two environments was the fact that access to the student performance records was immediate at any point for the IDLE system, while the WebCT users had to exit the quiz first, a tedious process that the students avoided to follow. This possibility however could have a negative effect on the IDLE students, since they can lose their concentration and be discouraged by possible poor score. At design time these subtle differences of the two environments in the navigation and interaction model seemed not crucial for the task and made the educators and designers believe that the usability and the performance of the students could not possibly be affected.

METHODOLOGY OF THE STUDY

Design of the Study

Following the framework described above, an evaluation experiment of the IDLE testing and self-evaluation module was performed. The objective of the experiment was twofold:

- To measure the usability and effectiveness of the module in comparison to the reference environment

- To investigate the impact of usability on student performance

Usability evaluation was performed through an on-line questionnaire that the students had to answer, immediately after they completed the main task. The task that the students had to perform during the experiment was to take an on-line test made of multiple-choice and fill-the-blank questions. The usability evaluation questionnaire was anonymous and was completed in a voluntary basis. From the 108 students that used the distance-learning software, 88 of them (81%) filled the usability evaluation questionnaire. The questionnaire, which is included in Appendix A, contained 10 closed questions with answers in a multipoint scale of five (5) values in the range: 1 (bad) to 5 (excellent), relating to key usability aspects. The questions are inspired by the heuristic evaluation rules of (Molich & Nielsen, 1990) that have subsequently been adapted (Nielsen, 1993) and widely used in the frame of heuristic evaluation experiments (Nielsen, 1992; 2000; Squires & Preece, 1999; Levi & Conrad, 1996).

The reason for which we opted for this evaluation technique is related to the following:

- It provides a quantitative measure of usability, i.e. it serves the objective of comparison of two alternative software environments and correlation to student performance
- The technique has been widely accepted as a concise test of usability and has been widely used
- This technique does not relate usability problems to their causes, thus it is not suitable for formative evaluation, however it is brief and therefore suitable for end-users, especially students, something that was confirmed in our case by the large number of students that filled the questionnaire.

The second objective was that of measuring the impact of the environment usability on student performance. This was measured by comparing the performance of the students that used the two alternative environments and relating it to software usability as measured by the

previous experiment. Also for reference reasons a part of the task was assigned to a small number of students who used traditional paper and pencil. Obtaining a quantitative measure of student performance was straightforward given the nature of the task. However the objective of measuring the impact of usability of the software on student performance presented difficulties, since there are many variables that might affect student performance and need to be controlled. So some effort was made to diminish the effect of other variables, like the conditions of software use and environment of interaction, the characteristics of the two student populations, etc., as described in the following section. In more detail a number of additional were performed: (a) an independent assessment of student performance in the subject in order to confirm the lack of bias in the formation of the two main user groups that participated in the experiment, (b) a study on the impact of the software environments on the task performance in comparison to a traditional paper and pencil environment, (c) a study on the impact of delays of software performance on task completion, (d) a study on whether the performance of each individual student has influenced his/her judgment on software usability.

Context of the Experiment

The experiment took place in the frame of a first semester University course of the Electrical and Computer Engineering Department of the University of Patras. This course (22Y103 Introduction to Computers I) involves an introductory laboratory part, during which students are introduced to the Computer Center of the Department and in particular to the Unix operating system and Internet theory and practice (HTML, email, ftp etc.). Distance learning material is provided to the students, supportive to the traditional laboratory teaching and experimentation. Halfway through this course during the Academic Year 1999-2000, the students were asked to prepare themselves for a number of diagnostic assessment tests on the

material covered in the frame of the laboratory. The tests were to be contacted during laboratory sessions. No indication was provided to the students on the means to be used for the tests. One hundred and twenty (120) students participated in the experiment that lasted three weeks. The students were divided in three groups in an arbitrary way. The first one of them, made of fifty-seven (57) students, used the IDLE software, the second one, made of fifty-one (51) students, the WebCT module and the third one, made of twelve (12) students used the paper and pencil environment. None of the students had previous access to the modules used, so no previous practice with the environments was assumed for any of them. The students had varying experience of use of computers and attitude towards use of computers in education. The material on which the students were tested was taught and studied in various ways, including distance-learning techniques. However it should be clear that the purpose of this particular experiment has been to measure the effectiveness of the testing and evaluation module of IDLE and not the overall effectiveness of the distance-learning environment. So no special inquiry was made on the use of the distance learning course material.

One of the most important factors, the influence of which had to be investigated, was the skill and performance of the students on this subject. So we collected data from the independently conducted final test on this Laboratory course, which was part of the standard educational process, at the end of the same semester. Then we examined the performance of the two groups of students that used IDLE and WebCT in this test. The results are shown in Table 1.

Table 1

Performance of the Two-student Groups in the Final Examination on the Subject

Student Group	*Average score*	*Standard deviation*	*Minimum score*	*Maximum. Score*
Users of IDLE	**3.95**	**1.75**	**1.0**	**6.7**
Users of WebCT	**4.09**	**1.55**	**1.0**	**7.0**

By performing a t test on the mean values of the two groups we obtained P(t)=0.7874, t=0.2714, considered no significant. So it was deduced that the random subject classification was not a threat to the internal validity of the collected data.

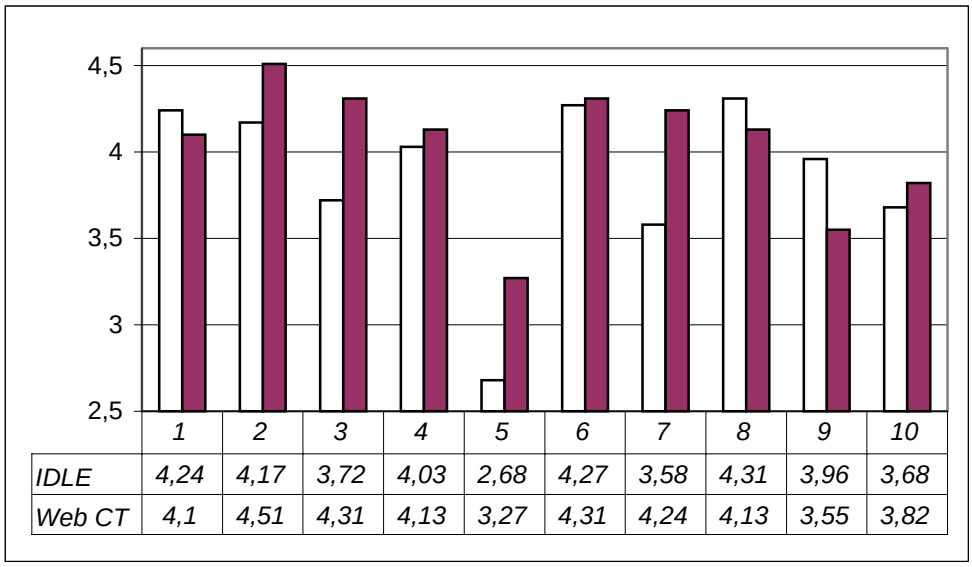
The conditions of the experiment were controlled, in order to eliminate the impact of any secondary variables on student performance. The students were first introduced to the subject. Each group used the software at the same time, so a significant but equal for all load was imposed on the server, thus simulating real distance-learning conditions. The students were located in the same computer room, so their behavior during the experiment was monitored. The time provided for doing the test was equal to all the students. The allocated time (approximately 30' for each session) was adequate for completing the test. However in case that a student requested additional time in order to complete unfinished parts of the questionnaire, this request was granted. Finally in order to eliminate the effect of possible delays of the network or the software on student performance, we did not take into account the not answered questions in students' evaluation, as discussed in relevant section below. The students were not informed about the last aspect; in contrary they were encouraged to complete all the test questions in the allocated time. They were also requested not to use any auxiliary material, like on-line help or handouts. The test was supervised, in order to establish that these rules were actually observed.

The students were informed that the test was going to have no effect on their evaluation for the subject, but it was going to have a diagnostic character as an indication of group performance. However it should be noticed, that the context of the test, i.e. in the frame of a laboratory session, together with the fact that the name of the student answering the test was known to the tutors made the students put a lot of effort in answering the quiz questions, as shown by their performance in the test, described in the following section.

There were three (3) sessions of 30' that took place during three consecutive weeks. During these sessions the students answered different sets of questions. The total number of questions was eighty-four (84). From them, thirty (30) questions were included in the first part of the test, thirty-two (32) in the second part and twenty-two (22) in the third part. Sixty-six (66) of the questions (78%) were of multiple-choice type with four alternative answers to each one of them and eighteen (18) were of the fill-the-blank type (22%). The subject of the three sets was the following: *1st session*: Simple Unix shell commands, *2nd session*: Advanced Unix shell commands, *3rd session*: Introduction to the Internet.

Usability Test

At the end of each session, the students, who had already used the Testing and self-evaluation module of IDLE or WebCT for considerable amount of time, were asked to complete electronically the usability evaluation questionnaire, included in Appendix A, for the module they used. At this point it was also explained that the replies to the questionnaire would have no impact on their score on the test or their final course grade. The usability evaluation was not compulsory and the students could submit their questionnaire anonymously if they wished to do so. An adequate amount of time was provided for completing the questionnaire. The students had the opportunity to go back to use the module, if they wished to do so. A number of open questions were also included at the end of the questionnaire, concerning their view on the usability and effectiveness of the module used. A considerable number of students filled the questionnaires: Forty–four (44) of those who used the IDLE environment and forty-four (44) who used WebCT. From them, forty-three (43), around 49%, filled their name in the questionnaire, while the rest submitted the questionnaire anonymously.



	1	2	3	4	5	6	7	8	9	10
IDLE	4,24	4,17	3,72	4,03	2,68	4,27	3,58	4,31	3,96	3,68
Web CT	4,1	4,51	4,31	4,13	3,27	4,31	4,24	4,13	3,55	3,82

Figure 3. Comparative presentation of evaluation of IDLE and WebCT .

RESULTS - DISCUSSION

The Usability Evaluation

The usability evaluation was performed in terms of absolute usability measure and, more significantly, as comparison of the two systems. The presentation of the results of the 88 questionnaires that were completed by the students is shown in figure 3.

The mean value of usability, if it is assumed that all ten questions were of equal importance, was 4.05 for WebCT and 3.82 for IDLE a difference of 5.7%, shown in Table 2. By performing a t test it is deduced $t=2.574$, $p<0.01$, considered significant. The overall result is therefore that while the students considered both systems of high usability, they thought that the WebCT environment was significantly more usable than the IDLE one.

Table 2

Overall Usability Evaluation

System	Number of	Average	sdev	Min	Max	Number of

	students	*score*				*questions*
IDLE	**44**	**3.82**	**0.41**	**2.83**	**3.91**	**440**
Web CT	**44**	**4.05**	**0.42**	**3.25**	**4.85**	**440**

Let us proceed with examining the details of the provided answers, shown in figure 3. From the table at the bottom of the figure, it can be seen that in three questions IDLE has a comparative advantage over WebCT (questions #1, #8, #9), while in the other seven the WebCT receives a higher score. The three questions for which IDLE has an advantage were those relating to system feedback (#1), aesthetic and minimalistic design (#8) and error recovery (#9). The advantage of the IDLE system about the feedback provided to the user (#1) is justified by the fact that, as mentioned in the previous section, IDLE provides feedback about correctness or not of a completed answer immediately after the submission, while WebCT does not provide such feedback. The aesthetic design (#8) of IDLE seems to have an appeal to the students over the simpler design of WebCT, while the error recovery issue (#9) is not very relevant in this context. From the other questions, one can observe that the one related to user control and freedom (#3) received relatively low score for both systems, probably due to lack of undo capability and limited freedom of movement around the test questions of the IDLE system. In relation to question about error prevention (#5), IDLE received a relatively low score, due perhaps to the inability of the system to prevent the user from selecting an already answered question. Finally in relation to flexibility and efficiency of use (#7) IDLE again scored poorly, since the cycle “select question, enter the question screen, answer the question, go back to the overview screen”, imposed to the user of IDLE, limits the efficiency of the system, considering in particular the delays relating to the described process.

Impact of Usability on Student Performance

As described in the previous section, the two main groups of students that participated in the experiment, using the two systems, had similar characteristics. An objective was to establish if the observed difference on usability between IDLE and WebCT, discussed in the previous section, had any impact on student performance. Student performance was determined from the scores in the assessment tests. In Table 3 the overall performance of the two groups is shown.

Table 3

The Student Performance for IDLE and WebCT

	Number of students	*Average score*	*Standard deviation*	*Minimum score*	*Maximum score*	*Number of questions*
IDLE	**57**	**6.65**	**1.211**	**3.3**	**9.0**	**1388**
Web CT	**51**	**7.34**	**0.866**	**5.5**	**9.7**	**1376**

The scores of the students were calculated as follows: For each correctly answered question, one point was given while for the incorrect ones no points were given. The final score was normalized in the range 1 to 10. The not answered questions were not taken into account. According to Table 3, the mean value of students' performance of WebCT was 7.34 and that of the students who used IDLE 6.65. So the WebCT users performed better than the IDLE ones.

The standard deviation of the WebCT users was 1.55, while that of the IDLE users was 1.75. By F test we obtained $F=1.955$, $p<0.001$ considered very significant. This difference in standard deviation is an indication that the WebCT environment is more stable and reliable as a tool for evaluation of student performance.

A test was performed in order to establish the statistical significance of the observed difference in performance. Because of the different standard deviations we performed a variation of hypothesis testing, which is the Welch corrected two-tailed unpaired t test. By

performing unpaired t test, which assumes that the two populations may have different standard deviations, we obtained $t(101)=3.42$ $p<0.001$ while the 95% confidence intervals were 0.2887 to 1.086, considered significant. The same observation holds for each one of the three laboratory sessions separately. In particular, for the fist session the twenty-one (21) students that used WebCT obtained an average score of 7.77, while the twenty-one (21) students that used IDLE scored 7.15. There was a statistically significant difference in students' performance ($t=2.245$ $p<0.02$). In the second session the eighteen (18) students that used WebCT scored in average 7.09 while the twenty-two (22) students that used IDLE scored 6,37. The unpaired two-tailed t test confirmed the statistically significant difference of the two values. ($t=2.184$, $p<0.02$ 95% confidence intervals 0.052 to 1.39). Finally in the third test, the twelve (12) students who used WebCT obtained 6.95 and the fourteen (14) students that used IDLE 6.34. The small population of the two groups in this case made us fail to reject the null hypothesis ($t=0,094$ $P(t)= 0.098$), so we cannot establish statistically significant difference in the performance of the two groups, however even in this case the trend of better performance of the WebCT users towards the IDLE users is maintained.

The validity of this significant finding of correlation of usability to student performance in this experiment had to be further checked according a number of dimensions as discussed in the three following sections.

Impact of the Electronic Environment on Task Performance

A test that took place concerned the impact of the electronic environments used on the assessment task. So additionally to the two groups of students that used WebCT and IDLE, a third group was formed that took the first examination using a paper and pencil environment. The comparative performance and the characteristics of the three groups that participated in this study are included in Table 4.

Table 4

Comparison of Paper & Pencil Environment to the Two Electronic ones

System	*Number of students*	*Average score*	*sdev*	*Min*	*Max*	*Number of questions*
IDLE	**57**	**6.65**	**1.21**	**3.3**	**9.0**	**1388**
Web CT	**51**	**7.34**	**0.87**	**5.5**	**9.7**	**1376**
Paper & pencil	**12**	**7.53**	**0.67**	**6.5**	**7.4**	**360**

A statistical analysis was conducted in order to check the significance in student performance variation of the three groups of Table 4. A non-parametric variation of ANOVA test was applied (Kruskal-Wallis test with Dunn post tests to check the difference among groups paired one to one). This test was used since a statistical significance between IDLE and WebCT groups' standard deviation of scores has already been deduced. (ANOVA requires no difference of standard deviation between groups). We obtained KW=10.456 (corrected for ties), $p<0.01$ considered significant (proof of variation existence). Further analysis with Dunn's multiple comparison test resulted in a significant existence of difference between IDLE and the other two groups: (IDLE -WebCT: mean rank difference (MRD=-17.202, $p<0.05$), IDLE - paper & pencil: MRD=-28.910, $p<0.05$, WebCT - paper & pencil: MRD=-11.708, $p>0.05$, not significant).

This analysis suggested that there is no significant difference in performance between the users of the paper & pencil environment and WebCT, while the difference between the paper & pencil environment and IDLE is significant.

One variable that seemed to be influenced by the tool used, was the time required by the various groups to complete the task. While the average time for the IDLE and the WebCT groups was 30 min, the average time of the paper and pencil (p&p) group was 22 min, thus 25% less. This is attributed to: (a) the processing and communication delays of the distance-learning environments, (b) the lower readability of text on CRT screens compared to printed

text of similar characteristics and (c) the unfamiliarity of the students with the new modules, in comparison to the familiar paper and pencil environment.

In conclusion, it is deduced from this part of the study that the most usable electronic environment (WebCT) was as effective as the traditional paper & pencil one, while it lacked behind in efficiency. In contrary the less usable electronic environment (IDLE) was not as effective or as efficient as the paper and pencil environment.

Study of Task Completion effect on Student Performance

One of our concerns was to establish possible other secondary parameters affecting student performance. One such parameter is related to delays due to the difference in client-server communication and implementation of the two environments. We would like to make sure that the observed significant performance variation between the two groups is not owed simply to the fact that due to lack of time, one of the two groups did not complete the test. This factor was eliminated by taking into account in the scores calculation only the answered questions. During the experiment the students were not aware of this fact, they were encouraged to answer all possible questions. In table 5 we include information concerning this aspect. As one can see in this table, the WebCT group of students answered 92% of their questions, while the IDLE group answered 83% of theirs. If the test scores had been based on the overall available questions the average performance of the WebCT group would have been 6,8 and that of IDLE students 5,6, thus making the difference in performance between the two groups even greater than that presented in a previous section (the impact of usability on student performance).

Table 5

Students' Performance in Relation to the Number of Answered Questions

System used	*Number of students*	*Session*	*answered questions*	*Score/ on answered questions*	*sdev*	*number of questions*	*score / on all questions*	*% of answered questions*
IDLE	21	s1	557	7.2	1.02	30	6.3	88%
	22	s2	542	6.4	1.22	32	4.9	77%
	14	s3	289	6.3	1.28	24	5.5	86%
	57		***1388***	***6.7***			***5.6***	***83%***
WebCT	21	s1	604	7.8	0.75	30	7.4	96%
	18	s2	494	7.1	0.74	32	6.1	86%
	12	s3	278	7.0	0.92	24	6.7	97%
	51		***1376***	***7.3***			***6.8***	***92%***
P&P	12	s1	360	7.5		30	7.5	100%

Correlation Between Individual Student Performance and Usability Evaluation

One aspect worth examining was to establish whether there is a correlation between the performance of individual students and their judgment over system usability. In other words, to examine if students who performed well in the test, thought that the system was more usable. Since the student performance is only partly related to usability and can be a result of other parameters, like skill, knowledge, practice, previous experience etc., such a strong correlation would have discredited in a certain extend the results of the usability evaluation experiment. This correlation has been studied by calculating Pearson r coefficient that takes values in the range 0 to 1, estimating the degree of correlation between two sets of values X and Y. This coefficient was calculated for the students who filled their name in the questionnaire. For the first session the value was r=0.25, while for the second one this value was r=0.02. Both values are considered low, indicating no correlation between the two data sets. An alternative way of examining this correlation is by depicting graphically the values in a scatter plot. An example of such diagram is shown in figure 4. By inspecting the diagram one can establish that there is no significant correlation between the two factors.

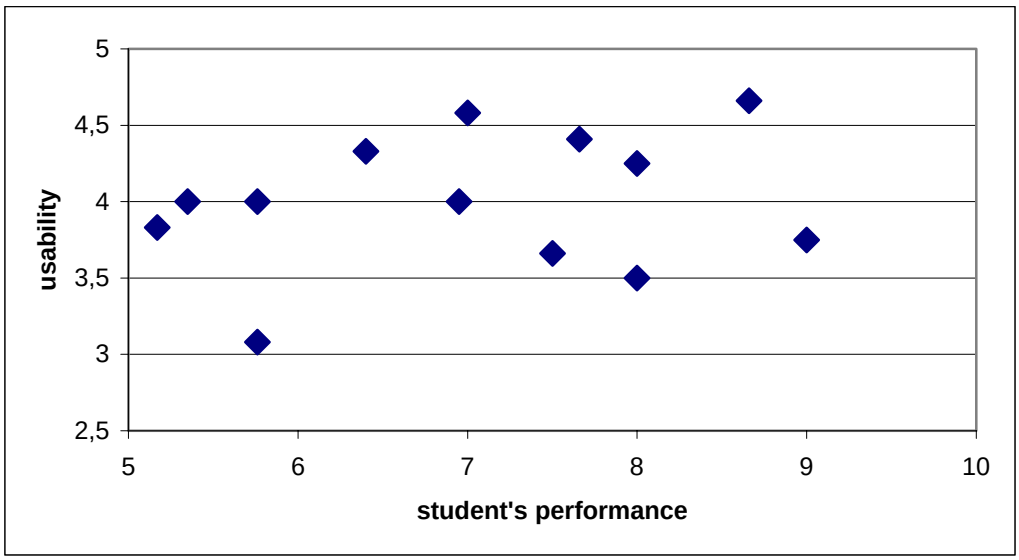


Figure 4. Correlation between usability and student performance (IDLE, 1st session)

CONCLUSIONS

The reported study concerned two complementary aspects of modern computer-support education and in particular distance-learning software: Usability evaluation of the systems and the impact of usability on the educational process. The usability of the software used was measured through the replies of the students who filled usability evaluation questionnaires while the student performance was measured by their scores in the assessment tests. While there is an ongoing discussion on the relation between usability and learnability (Squires & Preece, 1999) and there have been expressed objections on the relevance of usability in instructional software (Mayes, 1996; Jones, et al. 1999), from our research in the context of the reported experiment a correlation between the usability of the systems studied and the performance of the students in the studied task has clearly emerged. This correlation was evident in the results of the different sessions as well as in the results of the overall experiment. It seems that the most usable of the two systems had a positive impact on the performance of the group of students that used it. The reported difference in performance was statistically significant.

A considerable effort was made to create suitable experimental conditions in order to diminish the influence of other parameters on the two examined variables:

- The experiment was performed in controlled conditions (software, hardware and internet access) in order to eliminate the uncertainty of a typical distance learning situation
- All the students involved had no previous experience of use of the software modules involved
- The educational material used by the two student populations was identical
- The two main student groups had similar characteristics in terms of their background and performance in the subject, as confirmed by the results of the independent examination on the same subject at the end of the academic year, that took place a few weeks after the experiment

Both systems examined were characterized by high degree of usability, according to the usability evaluation test. In absolute values the IDLE module was evaluated as software of high usability, scoring 3.8 out of 5, using a widely accepted evaluation procedure. Many evaluation studies would have concluded at this point. However when IDLE was examined in comparison to a reference software (WebCT), a statistically significant difference in usability was measured. This is despite the fact that by inspecting the two modules the differences between them did not seem essential. Even more interesting was the finding that this measured difference in usability seemed to have a considerable effect on student performance. It seems that under the conditions that characterize the particular task, i.e. under time pressure the students involved had to understand the questions, reflect upon them, select the most appropriate question, thus performing demanding cognitive tasks, the subtle usability differences of the two environments played a significant role and thus had an impact on the effectiveness of the testing process. Additionally, the fact that the students were novice users

of the software modules under evaluation have increased the importance of software usability in this particular context of use. The usability in our study was related to quality of the software and in particular to efficiency in interaction, consistency, support in case of error, freedom in navigation, use of familiar to the user concepts. These are important issues in any educational context, since they permit the educational software to become transparent and not interfere with the learning process.

A general conclusion of this study relates to the importance of usability evaluation of educational software and in particular distance-learning environments. In spite of the fact that the module examined during this study was particularly simple and had many standard features, it was demonstrated that the usability of the system influenced considerably the educational process. In contrary to more traditional tools, the modern computer environments are less neutral since they seem to play inevitably a significant role in the educational process. Educators and developers of software tools, especially those of high degree of complexity, should therefore be concerned about determining this role and develop adequate techniques for diminishing any negative influence of the tool on the educational process. Also techniques that permit design of software tools with these characteristics should be defined. This objective becomes more difficult in cases when the task and context of use of the software is far more complex than the one discussed in this paper.

References

Alexander, S. (1995). Teaching and learning on the World Wide Web. In R. Debreceny & A. Ellis (Eds.). *Ausweb95: Innovation and Diversity*, (pp. 93-99). Ballina, New South Wales: Norsearch Limited.

Avouris, N., & Tselios, N., (1999). *The Infotronic Distance Learning Module, Contribution to the CBT* Kernel Final Report, Leonardo Programme, Project 96-1650, Patras.

Avouris, N.M., Tselios, N., & Tatakis E.C. (2001). Development and Evaluation of a Computer-based Laboratory teaching tool, *Journal of Computer Applications in Engineering Education,* Vol 9, (forthcoming).

Becker, D., & Dwyer, M. (1994). Using hypermedia to provide learner control, *J. of Educational Multimedia and Hypermedia*, 3(2), pp. 155-172.

Eklund, J., (1995). Cognitive models for structuring hypermedia and implications for learning from the World Wide Web. In R. Debreceny & A. Ellis (Ed.), *Ausweb95: Innovation and Diversity,* (pp. 111-117). Ballina, New South Wales: Norsearch Ltd.

Fitzelle, G., & Trochim W. (1996). Survey Evaluation of Web Site Instructional Technology: Does it Increase Student Learning? *Proc. of Annual Conference of the American Evaluation Association, Atlanta,* GA, November 1996.

Goldberg, M.W., Salari, S, & Swoboda, P. (1996). World Wide Web - Course Tool: An Environment for Building WWW-Based Courses, *5th Int. World Wide Web Conference*, Paris, France.

Hayes, R. (2000), Exploring Discount Usability Methods to Assess the Suitability of Online Course Delivery Products. The Internet and Higher education 2 (2-3), pp: 119-134.

Heines J.M. (2000). Evaluating the effect of a course web site on student performance, *J. of Computing in Higher Education,* 12 (1), pp. 57-83.

Jones, A., Scanlon, E., Tosunoglu, C., Morris, E., Ross, S., Butcher, P., & Greenberg J. (1999). Contexts for evaluating educational software, *Interacting with Computers*, 499-516.

Kirkpatrick**,** D. (1994). *Evaluating Training Programs*. San Francisco, CA: Berrett-Koehler Publishers, Inc.

Levi M.D., Conrad, A., & Frederick G. (1996). A Heuristic Evaluation of a World Wide Web Prototype, *Interactions Magazine*, July/August, Vol.III.4, pp. 50-61.

Mayes, J.T. & Fowler C.J. (1999). Learning technology and usability: a framework for understanding courseware, *Interacting with computers,* 485-497.

Mayes, T., (1996). Why learning is not just another kind of work, *Proc. on Usability and Educational Software design*, BCS HCI, London, December

Molich, R., Nielsen, J. (1990), Improving a human-computer dialogue, Communications of the ACM, 33 (3), pp. 338-348.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., & Arcuri, M. (1999). Comparative evaluation of usability tests. *Proceedings of ACM CHI'99 Conference on Human Factors in Computing Systems*, Panels, pp. 83-86.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings ACM CHI'92 Conference* (Monterey, CA, May 3-7): 373-380.

Nielsen, J. (1993). *Usability Engineering*, Academic Press, London.

Nielsen, J. (2000). *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, Indianapolis.

Squires D., & Preece, J. (1999). Predicting quality in educational software: Evaluating for learning, usability and the synergy between them, *Interacting with Computers*, 11, pp. 467-483.

Thuring, M., Mannemann, J., & Haake, J. (1995). Hypermedia and cognition: Designing for comprehension. *Communications of the ACM*, 38(8), 57-66.

Tognazzini B. (1992). *Tog on Interface*. Reading, MA: Addison-Wesley

Acknowledgements

The authors would like to thank Lefteris Fotopoulos who developed the CBT-Kernel application and participated in the reported experiment, Sophia Daskalaki of the Univ. Patras for comments on early draft of the paper and Prof. Costas Goutis for continuous encouragement and support. Financial support from the CBT Kernel/Leonardo project 96-1650, EPEAEK/ECE Curriculum Development Project and PENED 99ED234 Research Project are also acknowledged.

Appendix A. Heuristic Usability Evaluation Questions used in the study

(1) Does the system provide appropriate feedback about its current state within reasonable time?

(2) Is the language used by the system simple and comprehensible to you? Do you think that the visual and symbolic representations used at the interface are adapted to the intellectual level of the user?

(3) Do you think that the system provides you with adequate control and freedom of movement, for example support for undo?

(4) Is the system self-consistent in the use of terminology, semantics of symbols etc, across the user interface?

(5) Do you feel that the system protects the user from errors?

(6) Does the system require from the user to remember many things, does it make an effort to minimize user's mnemonic load?

(7) Does the system provide flexible shortcuts to experienced users needs?

(8) Is the system characterized by aesthetic and minimalist design so that it avoids irrelevant information that can create confusion to the user?

(9) Are error messages precise simple and constructive?

(10) Judge quality of provided help and handbooks