



May 19, 2001

WORKSHOP PAPERS

**Assessment Methods in
Web-Based Learning
Environments &
Adaptive Hypermedia**

Assessment Methods in Web-Based Learning Environments & Adaptive Hypermedia

May 19, 2001

Juan E. Gilbert
Auburn University
Computer Science and Software Engineering
107 Dunstan Hall
Auburn, AL 36849
gilbert@eng.auburn.edu

Roland Hübscher
Auburn University
Computer Science and Software Engineering
107 Dunstan Hall
Auburn, AL 36849
roland@eng.auburn.edu

Sadhana Puntambekar
University of Connecticut
School of Education
Educational Psychology Program in Instructional media and Technology
U-64, 249 Glenbrook Ave.
Storrs, CT 06269-2004
sadhana@uconnvm.uconn.edu

Contents

Introduction	1
<i>Juan E. Gilbert, Roland Hübscher and Sadhana Puntambekar</i>	
Self-assessment: how good are students at it?	2
<i>Antonija Mitrovic</i>	
Wired and Wireless e-Classroom Learning Assessment	9
<i>Jerome Eric Luczaj and Chia Y. Han</i>	
Making Sense of Open Representations Through a Constructive Hypertext: How to evaluate learning?	14
<i>Carlo Iacucci and Helen Pain</i>	
Layered Evaluation of Adaptive and Personalized Educational Applications and Services	21
<i>Charalampos Karagiannidis, Demetrios Sampson and Peter Brusilovsky</i>	

Introduction

Assessment Methods in Web-Based Learning Environments & Adaptive Hypermedia

As the number of web based learning environments such as adaptive hypermedia systems increases, the need for appropriate assessment methods increases as well. However, assessment is lacking in methodology and, for many existing systems, in empirical data. The objective of this workshop is to address the validity of assessment methods used in web based learning environments and adaptive hypermedia. The workshop also aims at developing new assessment strategies that can be researched for future systems.

When assessing web based learning environments and adaptive hypermedia systems, the distinction between evaluation and assessment must be explained. Evaluation refers to a system point of view where the system is at the center of the evaluation. Evaluation is concerned with system performance and the system's decision making capabilities. This may be specifically related to the adaptive decisions that the system makes, the efficiency and performance of the system and/or the algorithms that the system employs. In any case, the evaluation occurs from a system's point of view and conclusions about the effectiveness of the system are derived.

The term assessment refers to a learner centered approach to system evaluation. This approach attempts to evaluate the learner. In most cases, this is directly related to the assessment of learner outcomes. Learner outcomes are typically captured using some form of web based testing and/or some form of problem solving. The goal of learner centered assessment is to draw conclusions about the effectiveness of the system by measuring learner outcomes.

In both evaluation and assessment, the primary goal is to determine the effectiveness of the web based or adaptive hypermedia system. The web has been heavily populated with these systems that provide instruction and other services, but the effectiveness of these systems is questionable in most cases due to the lack of empirical data. By defining effective evaluation and assessment strategies, this workshop aims at providing a new means of measuring effectiveness in web based and adaptive hypermedia systems. This workshop hopes to create a working assessment group that will expose the research and development communities to the evaluation and assessment strategies discussed in the workshop and beyond the workshop.

Self-assessment: how good are students at it?

Antonija Mitrovic

Intelligent Computer Tutoring Group
Computer Science Department, University of Canterbury
Private Bag 4800, Christchurch, New Zealand
tanja@cosc.canterbury.ac.nz

Abstract: For effective learning, it is not enough to provide support for learning domain knowledge; it is also necessary to teach students how to learn. Various metacognitive skills are required for effective learning, and there are several recent projects investigating how to support the acquisition of these skills. This paper presents a study of whether students are able to critically assess their own knowledge. This particular metacognitive skill has been investigated in the context of SQL-Tutor, a system that helps students to learn a database language. We found that not all students are good at evaluating their own knowledge, and present several approaches to support students in learning this particular skill.

1 Introduction

Intelligent educational systems (IES) aim to provide individualized environments suited to the needs, abilities and knowledge of each student. If the goal is to maximize student's learning, it is necessary to teach the student how to learn effectively, not only to provide support for learning domain knowledge. Therefore, IESs must support students in developing metacognitive skills.

Metacognition has been studied in several disciplines, such as education, psychology and AI. It is generally accepted that metacognition includes the processes and activities involved with awareness of, reasoning and reflecting about, and controlling one's cognitive skills and processes. Metacognition therefore involves thinking about, inspecting and adjusting one's thinking, problem-solving approaches and learning habits, among others. Self [12] lists the following metacognitive activities: explaining something to oneself or others, being able to evaluate one's knowledge, planning, recognizing problems and their characteristics, allocating resources, applying appropriate strategies, monitoring and evaluating one's problem-solving approach, checking consistency of data etc. A number of studies showed that improved metacognitive skills result in improved problem solving and better learning [2, 5, 13], and that such skills can be taught [3].

Conati and Van Lehn have experimented with self-explanation, which is a skill of "generating explanations and justifications to oneself to clarify an example solution" [5]. In their case, the student is asked to explain a solution that is provided by the system. Aleven and Koedinger [2] explore how students explain their own solutions. In both cases, significant gains have been achieved by the students who explained the solutions. Reflection is encouraged by allowing the student to inspect and, in some cases, to modify the student model in [4, 6, 7, 11].

Aleven and Koedinger [1] evaluate students' abilities to identify situations when help is needed and to ask for appropriate help. They show that not all students possess this skill, and recommend several ways in which the system may support students in

acquiring it. They also point out that it is necessary for educational systems to model not only student's domain knowledge, but also his/her meta-knowledge. Such meta-model would enable the system to provide individualized support for learning at the meta-level as effectively as at the domain level.

This paper focuses on a different skill. If students are to learn, they need to be able to critically assess their knowledge. This skill is important in order to be able to identify what topic needs attention, which is necessary for problem selection. The same skill is also important for students to assess the difficulty of the problem they are working on, and to decide whether to abandon the problem or keep working on it.

The following section introduces SQL-Tutor, a system for teaching a database language, that is the context of this study. SQL-Tutor provides a facility for students to select problems on their own, which requires students to be able to evaluate their own knowledge. Section 3 describes the experiment performed in order to obtain the data, and is followed by a description of the findings in section 4. The conclusions are presented in the final section.

2 SQL-Tutor

SQL-Tutor is an intelligent educational system, which helps university-level students to learn SQL. The architecture of the stand-alone version of the system is illustrated in Figure 1. For a detailed discussion of the system, see [8, 9]; here we present only some of its features. SQL-Tutor consists of an interface, a pedagogical module, which determines the timing and content of pedagogical actions, and a student modeller (CBM), which analyzes student answers.

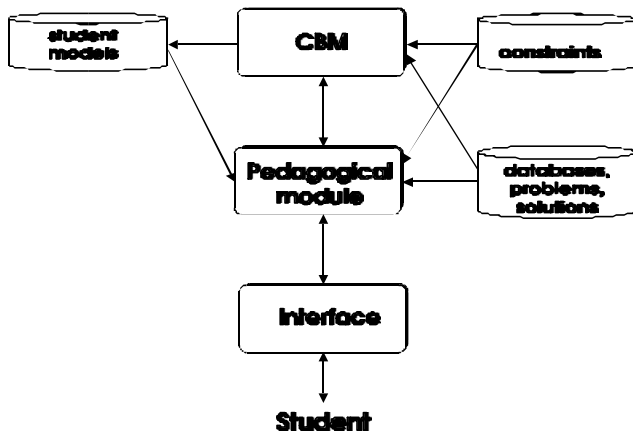


Fig. 1. Architecture of SQL-Tutor

It uses Constraint-Based Modeling [10] to model knowledge of its students. Students have several ways of selecting problems in SQL-Tutor. They may work their way through a series of problems for each database. The other option is a system-selected problem, when the system selects an appropriate problem for the student on the basis of his/her student model.

3 The Study

This project focuses on a student's ability to critically assess his or her own knowledge. Such ability is crucial for effective learning. Students should be aware of the extent and quality of their knowledge if they are to improve it. Reasoning about

one's knowledge is also necessary to be able to identify the gaps in the knowledge, and to select appropriate areas that require attention.

Assessing one's own knowledge is a difficult task. Our hypothesis is that students are not generally good in evaluating their knowledge. We propose that there are several factors that influence this ability. We assume that good students will be better when assessing their own knowledge than less able students. In other words, such ability depends on one's knowledge and experience. This is in accordance with findings from other studies [1]. We also assume that there are other personal factors that may be influential, such as the personality type etc. However, we do not propose to identify these personal characteristics, as they are not modelled in SQL-Tutor.

In order to evaluate our hypothesis, we performed an experiment on the SQL-Tutor system, which was modified slightly to allow for data collection. We focused on situations when students abandon the problem they are working on, and ask for a new problem. In such situations, the student was asked two questions. Firstly, we asked the student to specify the reason for abandoning the current problem. Three possible replies were offered: the student may think that the current problem is too easy or too difficult, or may simply want to work on a problem of a different nature. After specifying the answer to this question, the student is asked to specify what kind of problem they would like to work on next. For this purpose, problems were characterized by the clause, so seven options were available, one for each clause of the SELECT statement, plus the *any clause* option.

The students who participated in the study were enrolled in an introductory database course at the Computer Science Department at the University of Canterbury, New Zealand, in the second half of 2000. The usage of the system was voluntary. The system was demonstrated in a lecture at the beginning of September, and the students were told that SQL-Tutor is a good practice environment. The course involved a test on SQL a month and a half after the system was introduced. The experiment was set up this way so that the students may have use the system over several weeks. Prior to this experiment, SQL-Tutor has been evaluated in three studies in 1998 and 1999. All the previous evaluations of SQL-Tutor [8, 9] involved students using the system in a single, two-hour long session. Although we have gained a lot of experiences in these studies, we believe that longer studies are needed in order to see longer-term learning results. Hence, we decided to give the students the opportunity to study with the system whenever it suits them over a longer period.

Prior to the experiment, all students listened to two lectures on SQL. During the experiment, there were 5 additional lectures on SQL, and a series of five labs on defining and using databases in the context of the Oracle DBMS. The experiment required the student to sit a pre-test, which was administered as the first page when accessing the Web-enabled version of SQL-Tutor. The pre-test consisted of three multi-choice questions. The first and the last question were worth 1 mark each, as students were told that there is only one correct answer among those given. The second question was worth 5 marks, as there were 5 possible answers and students knew that there might be more than one correct solution. The questions were designed to test the student's knowledge of SQL.

After the pre-test, the students were free to select any of the databases and work on the problems. All students' actions were recorded in logs. The post-test was administered separately. It consisted of three multi-choice questions of similar nature to those in the pre-test. The post-test was administered on paper to all students enrolled in the course 7 weeks after the informal start of the experiment.

4 Results

This section presents the results of the analyses performed on the data collected in the experiment. Section 4.1 presents the general findings about how students learnt with SQL-Tutor. Data analyses relevant to our hypotheses are discussed in section 4.2.

4.1. Learning with SQL-Tutor

Out of 142 students enrolled in the course, 79 logged on to SQL-Tutor and sat the pre-test. However, some of these students have only briefly looked at the system. We excluded the logs of nine students who attempted no problems. Table 1 presents some simple statistics about the usage of the system.

The number of sessions that the students had with the system ranged from 1 to 7, with an average of 2. The length of sessions also varied greatly: the shortest session was only one minute long, while the longest took 300 minutes. The average duration of a session in the experiment was 47.45 minutes, and the average total time spent with the system per student was 95.6 minutes (the minimal total time was 2 minutes, and the maximal total time was 534 minutes). Since the interaction times were very different, it is not surprising that the number of problem attempted was also quite variable: the minimum number of problems attempted in a single session was just one, while the maximum was 30. As specified in Table 1, the average number of problems attempted in a single session was 6.65. The number of attempted problems that were eventually solved (per session) ranged from 1 to 27, with the average being 1.5. The total number of solved problems per student (during the total interaction time) ranged from 1 to 44, with an average of 10.26. The percentage of problems that were correctly solved ranged from 0% to 100%, with an average of 67.5%.

Table 1 also contains the results on the two tests administered. The maximum mark for the pre- and post-test was 7. The average mark rose from 4.02 on the pre-test to 5.01 on the post-test. The paired t-test was run on these results, and the difference between the pre- and post-test results is statistically significant ($t=-4.49$, $p=1.63E-05$).

Sessions	Session length (min)	Total time (min)	Problems Attempted/ session	Problems solved/ session	Total solved	Pre-test Mean (SD)	Post-test Mean (SD)
2	47.45	95.6	6.6	1.5	10.26	4.02 (1.52)	5.01 (1.24)

Table 1. Statistics about the interactions

4.2 Analyzing the self-assessment skills

The logs also contain the data relevant to our hypothesis. Out of 70 students, 25 had not abandoned any problems. The remaining 45 students abandoned at least one, and at most 15 problems, with an average of 3.87 abandoned problems per student. The total number of abandoned problems for all students was 165. The number of attempts before abandoning the problem ranged from 0 (the total of 98 cases) to 25 (an average of 2.49). Therefore, most often (in 59.39% of the cases) the students abandoned the current problem without making any attempts at it. The average number of problems abandoned after 0 attempts for students in this group was 2.24 (ranges from 0 to 12).

In order to evaluate our hypothesis, we divided 45 students from the experimental group who were of interest for our hypothesis into two subgroups, based on the results

of the pre-test. Students who scored above average (i.e., 5, 6 or 7 marks) on the pre-test were put into the *more able* group, while the students who scored 0 to 4 marks were put into the *less able* group. The groups were of similar sizes: 63.16% of students were classified as more able, as shown in Table 2. This table also presents the statistics given in table 1 for the two subgroups¹. The mean of the post-test was lower than the pre-test mean for the more able students, although not significantly. However, there was a drastic improvement in the means for the less able group. Therefore, such students benefited much more from working with the system than their more able peers. More able students tended to work longer with the system. The average numbers of problems abandoned after zero attempts were almost identical, but the more able students solved more problems.

Group	% of students	Pre-test	Post-test	Total time	Abandoned (0 attempts)	Solved problems
More able	63.16%	5.6 (0.75)	5.4 (0.94)	152.6	4 (2.3)	79%
Less able	57.4%	2.91 (1.06)	4.86 (1.49)	115	3.95 (2.31)	68.75%

Table 2. Statistics for the two groups of students with different prior knowledge

We now focus on what happened after the students asked for a new problem. Table 3 illustrates the situations encountered in regard to the problem that is obtained after abandoning the current problem with no attempts on it. The less able group was less likely to ask for a new problem without trying to solve the current one. The more able students were more successful at solving the next problem, and less likely to ask for yet another problem (the *abandoned* column). The rate of failure at the next problem was higher for the more able group, but close inspection of the logs revealed that more able students tended to work on more complex problems than the students in the other group.

	Success	Failure	Abandoned
More able	38.9%	11.1%	50%
Less able	35.5%	6.45%	58%

Table 3. The outcomes of the next problem obtained after 0 attempts

Next, we analysed the answers to the first question asked (what is the reason for abandoning the current problem). Out of the total of 165 abandoned problems, 57 (34.54%) were the problems from the more able group, and 108 (65.45%) were from the less able group. Therefore, less able students were much more likely to abandon a problem. The distribution of answers to this question is given in Figure 2. Less able students thought that the problem was too easy more often than more able students, although the inspection of the sessions very often contradicts the reason they specified. More able student asked for a different type of problems more frequently. Table 4 gives percentages of outcomes on the next problem following a specific answer to question 1. For each possible answer, the more able group was more likely to solve the next problem and less likely to fail at it than the less able group.

¹ Standard deviations are given in parentheses.

	Too easy			Too hard			Different type		
Able	Succ	Fail	No att ²	Succ	Fail	No att	Succ	Fail	No att
More	58.33	16.67	25	40	20	40	43.9	14.63	41.46
Less	24.49	26.53	48.98	7.69	69.23	23.08	32.35	20.59	47.058

Table 4. Outcomes following a particular answer to question 1

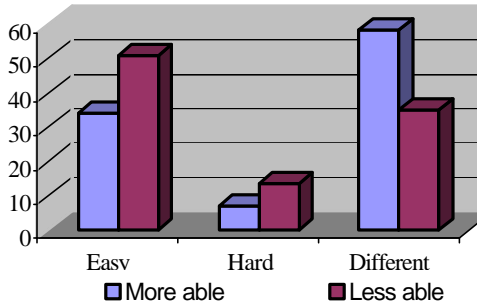


Fig. 2. Answers to question 1

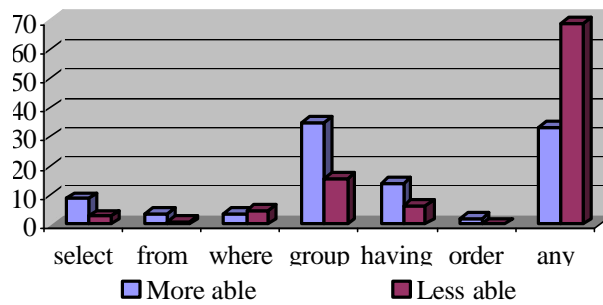


Fig. 3. Percentages of answers to question 2

An analysis of the second question (what kind of problem would the student like to work on next) showed that more able students were better at identifying the types of problems they needed to work on. Figure 3 shows the distribution of answers for the two groups. As we hypothesized, less able students were not good at identifying the kind of problem to work next, and therefore they specified *any clause* most often (in 69.44% of the cases). More able students asked for hard problems (*group by* and *having*) much more often than the other group (35.08% and 14.04% compared to 15.74% and 6.48%).

5 Conclusions

The study reported in this paper focuses on self-assessment as a metacognitive skill of high importance for learning. Students need to be able to evaluate their own knowledge in order to improve it. Self-assessment is important for problem selection, topic selection, selecting the appropriate type of feedback etc. Our hypothesis was that students are generally not good at evaluating their own knowledge, but that more able students would be better at this. We also expected less able students not to be able to select the appropriate type of the problem to work on next.

SQL-Tutor, an intelligent tutoring system for teaching the SQL database language, has been used as the context in which self-assessment was analysed. The study was performed with the second-year university students and concentrated on situations when students abandon the current problem. In such situations, we asked the students to specify one of three available reasons for giving up on a problem (too easy, too hard, or a request for a different kind of a problem). Following this, students were asked to choose the type of the next problem. There were six specific types that correspond to clauses of the SQL SELECT statement, and the *any clause* option.

The results of analysing the collected data justify our hypothesis. Majority of the students (64.3%) abandoned at least one problem. These students were divided into two groups, based on their mark in the pre-test. The analysis of their interactions revealed that less able students are much more likely to abandon a problem. These students are not able to select the type of the problem to work on next, and resort to

² *No att* means that the student has not tried to solve the problem at all.

the *any clause* option most often. More able students are much better at specifying the reason for abandoning a problem, and also better at selecting the type of problem to work on next. These students prefer harder problems (those that focus on the GROUP BY and HAVING clauses), which is consistent with our hypothesis.

An educational system must support students in acquiring metacognitive skills. As Aleven and Koedinger advise [1], the system should model not only students' knowledge, but also their metacognitive abilities. That way, the system may reason about the best way to support learning on the meta-level, not only at the domain level. There are several ways to improve the SQL-Tutor system so that it may support students in acquiring self-assessment skills. The system may intervene in situations when a student keeps abandoning problems without trying to solve them, and encourage the student to solve the problem. Also, the system could intervene when the student does not have a preference about the type of the problem to work on next. One way to help a student evaluate his/her own knowledge would be to visualize the student model. Since the student model in SQL-Tutor is quite complex, it could be summarized in a way similar to the answers offered for the second question. The student would then have a starting point to reason about their knowledge. Closer inspection of the student model may also have a positive effect on self-assessment skills. Future work on SQL-Tutor will involve the ideas presented in this paper.

Acknowledgements: The work presented here was supported by the University of Canterbury research grant U6430.

References

1. Aleven, V., Koedinger, K.: Limitations of Student Control: Do Students Know When They Need Help? In: Gauthier G., Frasson C., and VanLehn K. (eds): *Proc. 5th Int. Conf. ITS'2000*, Springer-Verlag, (2000) 292-303.
2. Aleven, V., Koedinger, K., Cross, K.: Tutoring Answer Explanation Fosters Learning with Understanding. In: Lajoie, S.P., Vivet, M. (eds): *Proc. Int. Conf. AIED (1999)* 199-206.
3. Bielaczyc, K., Pirolli, P., Brown, A.L.: Training in Self-Explanation and Self-Regulation Strategies: Investigating the Effects of Knowledge Acquisition Activities on Problem-solving. *Cognition and Instruction*, 13(2) (1993) 221-252.
4. Bull, S. See Yourself Write: a Simple Student Model to Make Students Think. In: Jameson, A., Paris, C., Tasso, C. (eds): *Proc. 6th Int. Conf. UM'97*, Springer, (1997) 315-326.
5. Conati, C., VanLehn, K.: Further Results from the Evaluation of an Intelligent Computer Tutor to Coach Self-Explanation. In: Gauthier G., Frasson C., and VanLehn K. (eds): *Proc. 5th Int. Conf. ITS'2000*, Springer-Verlag, (2000) 304-313.
6. Dimitrova, V., Self, J., Brna, P. (1999) The interactive maintenance of open learner models. In: S. Lajoie, M. Vivet (eds) *Proc. AIED-1999*, IOS Press, 405-412.
7. Kay, J. (1995) The UM toolkit for cooperative user modelling. *User Modelling and User-Adapted Interaction*, 4, 149-196.
8. Mitrovic, A., Hausler, K.: Porting SQL-Tutor to the Web. *Proc. ITS'2000 workshop on Adaptive and Intelligent Web-based Education Systems*, (2000) 37-44.
9. Mitrovic, A., Ohlsson, S.: Evaluation of a Constraint-based Tutor for a Database Language. *Int. J. on Artificial Intelligence in Education*, 10(3-4), (1999) 238-256.
10. Ohlsson, S.: Constraint-based student modeling. In: Greer, J.E., McCalla, G (eds): *Student modeling: the key to individualized knowledge-based instruction*, (1994) 167-189
11. Paiva, A., Self, J. (1995) TAGUS – a user and learner modelling workbench. *User Modeling and User-Adapted Interaction*, 4, 197-226.
12. Self, J.: Computational Mathematics (1995) <http://www.cbl.leeds.ac.uk/~jas/cm.html>
13. Swanson, H.L.: Influence of Metacognitive Knowledge and Aptitude on Problem Solving. *J. Educational Psychology*, 82 (1990) 306-314.

Wired and Wireless e-Classroom Learning Assessment

Jerome Eric Luczaj and Chia Y. Han
Department of Electrical & Computer Engineering and Computer Science
University of Cincinnati

Abstract: We discuss two major issues in e-Classroom learning assessment. First, we will discuss how to provide students an engaging environment that promotes active learning and second, how to evaluate learning outcomes with timely feedback in class.

1. Introduction

Any educational program has measurable goals used in assessment by accrediting organizations. Each program consists of a sequence of courses that deliver the core content of the discipline. Each course supports a portion of the program goals and its instructor is responsible to ensure that the course provides the appropriate content for the student. Learning outcomes are accomplished through course delivery and verified through test evaluation. At the end of the program, students should have covered all the required core knowledge and basic skills. However, there are many different factors that influence whether a student has successfully attained the learning outcomes. Often, the results are not homogeneous and the causes for poor student performance are hard to pinpoint. A fine-grained assessment strategy needs to be defined and implemented. Critical components of this strategy are a system for the instructors to demonstrate they have provided content supporting the program goals and a feedback system for the student to indicate that the content that the instructor thinks has been provided is the content the students believe they are receiving. In other words, the system must demonstrate that the content has been delivered and that it has been received.

2. Major Issues

Currently, course assessment is done in two distinct parts, student performance evaluation and teaching evaluation, and these two are not correlated. Student performance is evaluated at irregular intervals, through graded homework, quizzes, tests, projects, and final exams. Typically, students who were confused during lectures would not find out how far behind they were until it was too late to catch up. Although poor student performance, in general, is due to many factors, not solely related to the classroom experience, it is during the contact time in a classroom where the instructor can exert the greatest impact on student learning. In fact, a study from the University of Tennessee found that teacher effectiveness was the dominating factor affecting student academic gain.¹ Thus, it is important to assess learning in the classroom and let instructors take timely measures and make any necessary remedial changes.

In most cases, the main criterion for assessing student performance is the degree of subject matter understanding. However, from a holistic point of view, human factors should also be considered.^{2 3} Secondary factors that may significantly influence the student learning experience could include: study habits, including preparation for tests; understanding of pre-requisites material; attitude toward class, for instance, individual interest in subject matter or attentiveness to class progress; attitude toward instructor and compatibility between student learning style and instructor presentation style; attitude toward learning. These secondary factors impact how well students understand the core concepts of a course and how many of the learning outcomes they will achieve.

In general, at the collegiate level, the burden of teacher and course evaluation falls upon the student using either in-class or Web-based survey forms. Typically, these evaluations are done just once, if at all. Since the survey is normally completed toward the end of the academic term, it does not impact students in the current class, though it may be helpful to future students in the same course with the same instructor. Further, since students may not take the survey seriously, the validity of the student response is questionable. Additionally, these surveys often serve many purposes making their results imprecise. For example, a typical evaluation survey would cover a wide range of aspects, such as: course materials, including syllabus, reference material, textbooks, and hyper-linked sites; evaluation method relevance, fairness, timeliness; instructor presentation skills and attitude; facilities; course content.

Educational assessment has multiple, distinct uses in instructional improvement including: school and student accountability for academic achievement, feedback for teachers to revise teaching and administrators to allocate resources, and stimulation for deeper student understanding.⁴ The important thing to notice regarding the usual practices is that there is no correlation between student learning assessment and instructor teaching evaluation. Without timely feedback connecting student learning to instructor evaluation, neither the instructor nor the students have the power to affect change or to correct problems.

3. Proposed Methodology

With technology, it is now possible to enhance learning through active interaction between students, the instructor and the material. Within electronic classroom environments (e-classrooms), web-based instruction and networked computers or ubiquitous computing/PDA-based terminals are available for students to actively engage in note taking and providing instructor feedback.

We will first present a new framework for classroom and student achievement assessment where the followings are possible: make the purpose, expectations, and presentation explicit; engage in more frequent questions to evaluate and promote learning; collect timely feedback. Then, we will consider the main issues of assessing teaching and learning.

3.1. Brief Overview of CaSA

CaSA (Classroom and Student Achievement assessment) is a flexible framework for achievement assessment and student learning. The intent is for CaSA to augment the classroom experience by coordinating and synchronizing instructional streams, matching class plans to student class experience, and presenting instruction in a variety of media forms to promote self-directed learning. The emphasis is on facilitating timely feedback from students and offering alternatives to students with differing learning styles.

CaSA will consist of and coordinate three major components (see Figure 1). The Preparation component allows instructors or course administrators to define the e-classroom experience and Course Objectives. The Real-Time Stream component synchronizes instructional streams with student notes, assessment and feedback. The Review component provides the ability to review and evaluate instructional streams as well as feedback and assessment data gathered by the Real-Time Stream component.

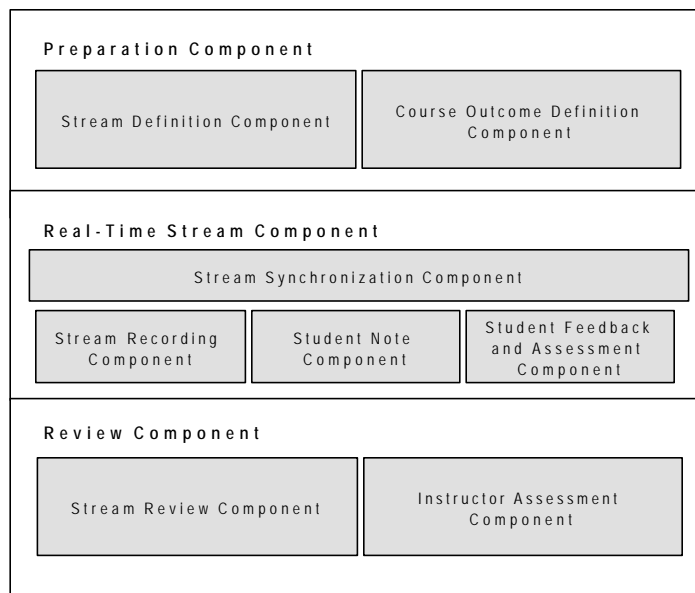


Figure 1: CaSA Component Diagram

Currently, a typical classroom experience involves multiple, simultaneous activities or streams of information. These may include, but are not limited to, verbal and visual instructor presentations, board notes, electronic exercises, overhead presentations, hyper-linked multimedia reference sources, student questions and notes, as well as student feedback and assessment. CaSA will synchronize these streams, providing the context and valuable insight during assessment, instructor feedback, and student review.

3.2. Current capabilities of the new e-Classrooms on UC campus.

Many of the classrooms on the University of Cincinnati (UC) campus are being converted and equipped with the latest e-media technology by the UC Instructional Media Center and Stremedia, the UC Streaming Media Project. The UC Stremedia Mobile Unit for in-class media acquisition and streaming data encoding accepts up to four input channels from digital pan/tilt based camcorders. These streams are integrated with digital cameras and Elmo cameras frame captures, Powerpoint presentation slide projection, and Mimio whiteboard inputs with post-production facilities. These digitized lecture streams are made available on the web for student review.

All UC College of Engineering students, beginning their sophomore year, should have their own computers. The college recommends students buy portable computer with wireless connection capability. Currently, three Engineering buildings offer wireless connections. Through the wireless connection in classrooms, student will be able to access the CaSA framework and NotePad control application. As such, these machines will present an excellent opportunity for immediate student feedback and assessment.

3.3. Assessment Strategy

Each lecture supports learning objectives from the course syllabus. Prior to the class, the instructor defines the lecture outline and the lecture notes (ClassPlan) which are associated with the text and course syllabus. The lecture consists of a ClassPlan presentation. As it is delivered, the various streams of presentation will be segmented into chunks, called PresentationBits (PB), each of which is associated with a media type and descriptive labels, such as 'introduction,' 'concept,' 'equation,' 'illustrative example,' etc. The flow of the lecture as well as the relationship of these PB entities should be described in terms of time-based media players (MP). With all the terminologies defined, the course plan can be represented in an XML-based format file.

The assessment strategy will take three forms: student topic marking, periodic and frequent electronic concept questions, and instructional stream review and evaluation.

First, the instructor ClassPlan will be made available to students on their computing device. As the class presentation proceeds, a student will be able to mark when a specific topic is covered. For the student, these marks will individually index the instructional streams. When they wish to review course material, the student can access the web-delivered instructional streams based upon their individual marks. By allowing the students to use their individual marks to access course material, this system provides an incentive for student participation. For the instructor and program assessors, the collected marks provides immediate feedback on when and if a topic has been successfully communicated to the students.

Second, the installed network of computers will permit the instructor to ask frequent, periodic questions to assess student understanding of the main concepts from the ClassPlan. Answers can be collected and immediate feedback regarding the overall state of student understanding can be presented to the instructor. In addition, reaction and

general attitude toward the material and class can be surveyed through positive and constructive questions during and at the end of the class.

Concept questions should provide supportive and corrective feedback, reinforcing correctly answered questions and giving correct answers with an explanation to incorrectly answered questions. Also, these questions should be integrated with an intelligent FAQ data bank to promote self-directed learning.

Third, by reviewing their presentation, instructors will be able to critique and improve their material delivery. Also, independent evaluators can assess a course (and by extension a program) by sampling instructional streams to ensure that all the material necessary to support the course learning objectives was covered.

4. Concluding Remarks

E-classrooms provide a unique opportunity to augment the current classroom experience with timely assessment, feedback, and information capture and review. By presenting lectures in a variety of media forms, students will be able to choose and access the particular streams that best suit their learning styles and self-direct their learning.

Given a framework where both instructional and presentational aspects of a lecture can be evaluated expeditiously, instructors and students can react quickly when there is a gap between intent and understanding. Coordinating various instructional streams with student assessment and feedback will provide the means for instructors to know when and if their intended message was communicated to their students. Also, instructors, by reviewing their presentations, will be able to critique and improve their delivery.

E-classrooms with appropriate frameworks will provide assessment data needed by program and course assessment staff to demonstrate course support of program goals. By developing a flexible instructional infrastructure, a bridge between course objectives and course assessment, classroom instruction and student feedback, and seat-time and study-time can be provided.

References

¹ Sanders, William L.; Wright, S. Paul; Horn, Sandra P., "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation," *Journal of Personnel Evaluation in Education* Volume: 11, Issue: 1, April 1997, pp. 57-67

² Mehrabian, A., *Silent messages: Implicit communication of emotions and attitudes* (2nd Edition). Belmont, CA: Wadsworth (1981)

³ Gorham, J., "The relationship between verbal teacher immediacy behaviors and learning: Monitoring processes and product" *Communication Education*, 43(4), 27-8, 1988

⁴ Baker, E. L.; Mayer, R. E., "Computer-based assessment of problem solving," *Computers in Human Behavior*, 15, pp. 269-282 (1999)

Making Sense of Open Representations Through a Constructive Hypertext: How to evaluate learning?

Carlo Iacucci and Helen Pain

Division of Informatics, University of Edinburgh, Scotland

{carlo,i,helen}@dai.ed.ac.uk

Abstract

This paper addresses the evaluation of educational tasks that make use of hypermedia in which representations characterised by a degree of openness are displayed. Specific problems of evaluating the educational practices are addressed, such as how do we deal with the openness of the representations in an open-ended task.

The case in which the representations include video clips of tutorial dialogues is considered. We discuss how evaluation should be approached in order to support the design of a specific kind of educational practice, namely, one in which learners are engaged as scriptors and use constructive hypertexts for reconciling different representation modalities and build coherent multimedia presentations (documentaries).

1. Introduction

When evaluating hypertext use for educational purposes, we are not only concerned with differing perceptions of worth and multiple dimensions of quality of learning experiences. There are also problems in observing hypertext users' behaviours (Pang and Edmonds, 1999). These problems are relevant in as much as one relies on a method based on defining educational objectives in terms of changes in the behaviour patterns of learners. However, the application we are considering relies heavily upon the *open* nature of the representations and we also have problems because we must deal with such openness in an open-ended task. In addition, defining the educational task must be part of our research. To accommodate these problems, we shall reconsider the relationship between evaluating and defining the educational objectives.

Open texts in hypermedia A common problem in hypermedia design and evaluation is how to deal with representations characterised by a degree of openness, that is, representations which lend themselves to alternative readings (open texts). In educational applications, such representations can either be used as (a) framed representations 'provided' with closure, which drive readers towards a 'preferred reading', or (b) resources to be accessed and freely interpreted in more user-centred tasks. Section 2 will address evaluation issues for case (a). We will consider case (b) here and will explain how we intend to rely on the property of open texts that invite readers (users) to participate. In the latter case, in the interaction set by the creator of an open text, audience's 'correct' answers are less important than the possibility of a *proactive* response.

As an instance of open texts, we will be concerned with spontaneous or scripted tutorial conversations. Each video clip is considered to be a single 'dialogue episode'.

The educational value of learning by observing others' dialogues The particular hypermedia application we are concerned with exploits the way students learn by observing tutorial dialogues. For the purposes of our discussion, we refer to results of research on observational learning from others' dialogues.

By showing videorecordings of verbal interactions between two learners or between a tutor and a learner one can obtain a comparable learning outcome to that of showing explanatory discourse of a teacher (Cox et al., 1998). So far, learners' spectatorship of such dialogues has been studied as a stand-alone activity, as opposed to being task-oriented and situated in the context of a more complex educational practice.

Using videos of pairs of students engaged in problem solving sessions, we carried out an analysis of learners' spectatorship, relying on the utterances through which spectators commented on the video (Iacucci and Pain, 2000). A synchronous account of the semantics of both spectators' comments and the dialogue context of the observed dialogue was used to characterise the indeterminacy of the problem. Limitations of traditional approaches to multimedia design which seek to minimise cognitive load and achieve semantic congruence in the presentations were identified.

As a consequence of these results, and through further empirical studies, we have increasingly become aware of the need for an analysis which takes into consideration the activities in which learners' are engaged both in producing and while observing such videos.

Several properties of these representations of dialogic interactions have been singled out (McKendree et al., 1998). We have looked at how such videorecordings can motivate and drive tutoring activities within a group of learners. They can provide several points of entry to a problem, or different viewpoints to the same knowledge domain. They can constitute representations to be accessed, edited, commented on and used to carry out constructive activities (see section 3).

Related research on Evaluation Traditionally, evaluating is concerned primarily with defining objectives (Guba and Lincoln, 1989). But this has not always been the case, as some methods for evaluating interactive applications in general and hypermedia in particular have partly freed the evaluation task from the need for identifying precise objectives (e.g. Nielsen, 1994; Garzotto and Matera, 1997). Such methods consider multiple values and allow the evaluator to rely also on her "connoisseurship" qualities to express 'judgements'. The approach of expert evaluation heuristics developed by Nielsen provides usability heuristics and it has been recently reconsidered in a 'crossed assessment framework' for linking educational goals (Squires and Preece, 1999). Squires and Preece integrate usability features with principled educational design by recasting Nielsen's HCI evaluation paradigm in terms of socio-constructivism. What is most relevant for our perspective is the consideration of aspects relevant to a socio-constructivist perspective of learning, such as intrinsic feedback, multiple views/representations, complexity and pedagogical techniques for constructivist approaches (scaffolding, bridging, anchoring, problem based environment and cognitive conflict) (Squires and Preece, 1999, p 477).

There are still difficulties in applying such approaches because of the openness of the multimedia representations we are concerned with and because of the openness of the end of the task through which such representations are accessed and re-used. Following (Mayes and Fowler, 1999), a "focus on the design significance of the nature of the

educational task” should be maintained (op. cit., p. 488). In our case, a difficult matter is identifying the very nature of the educational task. Evaluation should play an important role in such an endeavour through a research methodology in which evaluation is tightly bound to the creative aspect of the research work, as in action research (Argyris et al., 1987). The next section will look at what design commitments must be made when producing and framing the open representations for the hypermedia application and what implications this has for evaluation.

2. Open representations ‘framed’ in hypermedia: how can designers evaluate hypertext users?

When we first considered the case of explorative hypertexts, of which either teacher or learners were designers, a major conclusion we drew from our attempt to evaluate such educational experiences was that we should consider the representations as artifacts. This section will describe some limitations of evaluating by ‘matching’ appropriate content to appropriate learners in an appropriate context. We will argue for a more activity-based evaluation, as indicated in section 4.

When they are framed in such hypertexts, even if they display segments of ‘spontaneous’ conversations, dialogue episodes should be treated as artifacts, as opposed to as ‘mimesis’ of reality. In fact, designers must intentionally decide on a number of aspects which influence their access and readership:

- (1) *Deciding topics.* Even when episodes display segments of a spontaneous conversation, the designer acts as a selector.
- (2) *Intentions: communicative functions* can be assigned to such dialogue episodes. They might be intended to convey a ‘message’ which might not be in their text.
- (3) *Choosing a target audience* For instance, different audiences can be identified according to different stages of understanding of the issues.
- (4) *Selecting a representation format.* (choices on the length, the presence of explicit links, transcripts).
- (5) *Framing:* the dialogue episodes’ effect on a audience depends also on their title, on how they are introduced, annotated or provided with a follow-up explanation.
- (6) *Determining the observer’s context of spectatorship:* by motivating an inquiry, or creating a purpose for spectatorship.
- (7) *Allowing for interaction with the representation:* include annotation and linking, other than basic accessing, rewinding and forwarding facilities.
- (8) *Parsing the representation:* for instance, how to provide addressability of ‘sub-components’ of the videos (Faraday and Sutcliffe, 1997; Lee Tiernan and Grudin, 2001).

Since the audience was not given any goal to achieve through observing the dialogues, the evaluation of the process of spectatorship was addressed through the assessment of separate aspects. The method is best represented through a *sequence* of tests: how many individuals in the audience belonged to the target population? How many of those who belonged to the target population found themselves in the right context when observing the dialogues? How many of the remaining observers effectively perceived the content? How many of those who satisfied all the previous criteria understood what they

perceived? For how many of these remaining ones was the ‘desired effect’ achieved?

Similar ideas of ‘matching’ the right learners with the right representations in the right context have been considered in an application similar to ours:

“The challenge, of course, is to design a way in which appropriate dialogues can be matched to the learners immediate learning need.” (Mayes and Fowler, 1999, p. 492)

Problems This approach to evaluation causes problems. Some are related to the degree of openness that dialogue episodes retain in the hypertext. Others are related to hypertext exploration, to associative linking and intertextuality.

Following Bakhtin, texts’ meaning is created in relation to other texts, through *intertextual* relationships (Worton and Still, 1990). This becomes critical in hypertexts because both hypertext’s users and designers exercise associative linking. The links, be they implicit in the reader’s interpretation, or explicitly implemented in the hypertext, can constitute intertextual relationships of different kinds (Genette, 1997). As consequences of these phenomena in our case, we can name four reasons why they affect the evaluation of our educational practice:

1. *There is no original expository project.* The dialogue episodes have not been produced as a hierarchically structured tutorial exposition.
2. *Multiple ‘points of entry’ to a topic are provided,* as the content is meant to be displayed through a non-linear exposition.
3. *Often there are alternative interpretations to the preferred reading.* The dialogue videos are characterised by a degree of *openness*.
4. *Transience of the representations.* This posed difficulties for evaluation, as learners’ attention is selective (Faraday and Sutcliffe, 1997).

In order to overcome the difficulties, we focussed on the activities performed by learners. We will next consider the educational practice and activities, and the aspects evaluated, together with further research questions and goals.

3. Learners as scriptors of constructive hypertexts: the “documentary exercise”

Learners have been engaged in designing multimedia presentations by retrieving the dialogue episodes and framing and linking them to either segments of the on-line lecture notes or to their own scripts. In performing such activities, learners first access multimedia resources in an initial configuration. Next, they choose a perspective, for example, by identifying a question to be motivated and addressed. Then, they rearrange the objects to constitute a coherent exposition.

In this educational practice we are considering learners are engaged as ‘scriptors’. The kind of scripting they perform involves selecting, combining and framing (introducing and concluding) representations of different modalities. This kind of activity has also been termed “compilation scripting” and resembles the activity of creating a documentary by rearranging an archive of naturalistic documents (Ulmer, 1989). The directions given to the learners for carrying out the task are vague enough for the task to be considered open-ended. However, the artifacts produced, like TV documentaries, are required to provide a linear perspective, and to be characterised by closure, consistency and completeness.

This approach differs from other current applications that foster asynchronous learning through multimedia annotation (e.g. Lee Tiernan and Grudin, 2001). In our case learners are engaged as scriptors in constructing a compound artifact.

Constructive hypertexts Such activities can be devised through the use of ‘constructive’ hypertexts. The intention is to enable learners to use a full range of cognitive skills to focus upon the discovery of coherent structures and linkages. We take the term ‘constructive hypertexts’ as it has been used by Joyce:

“Learning is multiple yet integrative, difficult yet universal, not easily schematized yet apparently systematic, inherently personal and yet socially manifested, and so on. These contraries provide cautionary measures against which to judge exploratory hypertexts as learning tools. They also introduce and outline the promise of what I have termed constructive hypertexts. Every well-designed exploratory hypertext proceeds from a constructive hypertext created by its author or team of authors. ... The authors and audience of hypertexts share a transforming interrelationship. They are, to use an overused term, *colearners*.” (Joyce, 1995, p. 44)

Consequences for Evaluation Hence, learners have authored hypertexts, they have abandoned their status of ‘hypertext audience’ and participated as designers. Learners-designers can be confused with teachers-designers. This is a very problem of evaluating such practices: in such constructive activities learners are engaged in being teachers of the meaning they have created by structuring the representations. Evaluating their learning experiences involves evaluating them as designers.

4. Open representations in open-ended tasks: general aspects of evaluating the activity of designing-users

What is the consequence of considering the hypertexts as artifacts of which learners are authors? As we mentioned above, some difficulties in evaluating arise because by the interaction with the system, every reading by every reader becomes privileged, and authorised. In this section we indicate how we intend to support the constructive aspect of the practice and take into account the specific problems and opportunities mentioned above.

Evaluation in action research The educational objectives must be discovered while performing the practice with real learners: to do this forces the creative endeavour of the research to be carried out by observing the effects of introducing change into an educational practice. Hence, the building phase of the work and the evaluation phase belong to the same methodological step. This characteristic is part of the definition of *action research* (e.g. Argyris et al., 1987) and of methodological accounts which have been given to define *applied AIED* (Conlon and Pain, 1996).

In order to better motivate this point, we can state the following three aims of the practice, and the consequent goals for evaluation:

1. *The real objective is ‘to enable’* The system we are concerned with (the constructive hypertexts and the educational practice) cannot be held entirely accountable for what users actually learn or fail to learn. The complexities of the educational task, together

with other factors related to the specific media, and openness of the representations do not guarantee the learning outcomes. We should better see the system as an enabler. As in other constructivist approaches, the system should be accountable for how it can provide the best possible opportunities for learning to take place, by providing adequate opportunities for the user to challenge her own attitudes (Honebein et al., 1993).

2. *Learners' control and self-directed learning* The learner approaching the practice may not be able to formulate the most productive perspective and may need considerable help. But the system itself has to create opportunities for learning while not cauterising learners with its responses.

3. *Emergence of the educational objectives* A lot of what goes on during the educational practice we set is unforeseen. The educational objectives are emergent. Although an educational programme needs clear initial aims, in this open-ended educational practice the aims may be modified in view of learners' actual responses or even changed altogether. In fact, the actual effects of the practice may be more important than those intended. Furthermore, the practice must be adapted to meet the demands of different groups.

A sound approach to evaluation in the educational practice we have been considering should make such aims more visible. Following points 1 and 2, evaluation should provide insights on how many opportunities for further enquiry the educational practice creates for learners. Consequently, we should seek an account of the accessibility of the representations' content, and what potential interpretations are favoured from the learners' perspective.

Evaluation naturally induces focus on objectives, as it aims to give a perception of worth. But, following point 3, because of the nature of the task we should seek to discover and adapt objectives and methods while applying the practices.

Acknowledgements Thanks to Paola Baruchelli, her colleagues in the Net Quality Project at the Laboratory of Information and Communication Technologies of the University of Trento and to Giulio Iacucci for their contribution in carrying out the empirical investigations for the 'documentary exercise'.

References

- C. Argyris, R. Putnam, and D. Smith. *Action Science*, Jossey-Bass, San Francisco, 1987.
- T. Conlon and H. Pain. Persistent collaboration: a methodology for applied AIED. *Journal of Artificial Intelligence in Education*, Vol 7, No 3, 219-252, 1996.
- R. Cox, J. McKendree, R. Tobin, J. Lee, and T. Mayes. Vicarious learning from dialogue and discourse. *Instructional Science*, **27**, 431-458. 1998.
- P.M. Faraday and A.G. Sutcliffe. Designing effective multimedia presentations. *ACM CHI 97*, 272-279, 1997.

- F. Garzotto and M. Matera. A systematic method for hypermedia usability inspection. *The New Review of Hypermedia and Multimedia*, Vol. 3, 39-65, 1997.
- G. Genette. *Paratexts: Thresholds of Interpretation*, Cambridge University Press, 1997.
- E.G. Guba and Y.S. Lincoln. *Fourth Generation Evaluation*, Sage Publications, Newbury Park, 1989.
- P.C. Honebein, T.M. Duffy, and B.J. Fishman. Constructivism and the design of authentic learning environments: context and authentic activities for learning. In *Designing Environments for Constructive Learning*, Springer-Verlag, 87-108, 1993.
- C. Iacucci and H. Pain. A structured view of dialogue context as a basis for addressing the interactive re-use of educational dialogues. *Building Dialogue Systems for Tutorial Applications*, AAAI Symposium, November 3-5 2000.
- M. Joyce. *Of Two Minds. Hypertext Pedagogy and Poetics*, University of Michigan Press, 1995.
- J.T. Mayes and C.J. Fowler. Learning Technology and Usability: a framework for understanding courseware. *Interacting with Computers*, **11**, 485-497, 1999.
- J. McKendree, K. Stenning, T. Mayes, J. Lee, and R. Cox. Why observing a dialogue may benefit learning. *Journal of Computer Assisted Learning*, **14**, 110-119, 1998.
- J. Nielsen. Usability inspection methods. In J. Nielsen and R.L. Mack, editors, *Usability Inspection Methods*, John Wiley, New York, 1994.
- K.W. Pang and E.A. Edmonds. Modelling the Learner in a World Wide Web Guided Discovery Hypertext Learning Environment. In M.A. Sasse and C. Johnson, (eds), *Proceedings of INTERACT'99*, IOL Press, 251-265, 1999.
- D. Squires and J. Preece. Predicting quality in educational software: Evaluating for learning, usability and synergy between them. *Interacting with Computers*, **11**, 467-483, 1999.
- S. Lee Tiernan and J. Grudin. Fostering engagement in asynchronous learning through collaborative multimedia annotation. *INTERACT'01*, 2001.
- G. Ulmer. *Teletheory: Grammatology in the Age of Video*, Routledge, 1989.
- M. Morton and J. Still (eds). *Intertextuality: Theories and Practices*, Manchester University Press, 1990.

Layered Evaluation of Adaptive and Personalized Educational Applications and Services

Charalampos Karagiannidis and Demetrios Sampson

Peter Brusilovsky

Informatics and Telematics Institute (I.T.I.)
Centre for Research and Technology - Hellas (CE.R.T.H.)
1, Kyvernidou Street
Thessaloniki, GR-54639 Greece
Tel: +30-31-868324, 868785, 868580, internal 105
Tel: +30-31-868324, 868785, 868580, internal 213
E-mail: karagian@iti.gr, sampson@iti.gr

School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
Tel: +1-412-624-9404
Fax: +1-412-624-2788
E-mail: peterb@mail.sis.pitt.edu

Abstract. In this paper we address the evaluation for adaptive and personalized educational applications and services. We advocate the employment of layered evaluation, where the success of adaptation is addressed separately at two distinct layers: interaction assessment, and adaptation decision making. We demonstrate the benefits of this approach through two specific examples: we outline how layered evaluation can improve the evaluation of the KOD system; and how a previous study of the InterBook system can be revisited in the light of layered evaluation.

1 Introduction

Adaptive educational applications and services have attracted considerable interest worldwide, due to their potential to facilitate personalized learning, i.e. adapting to the individual requirements and preferences. Related literature includes a number of R&D efforts in the areas of intelligent tutoring systems, adaptive educational hypermedia and hypertext, intelligent pedagogical agents, etc. These efforts have resulted in a number of fruitful systems, which are differentiated according to a number of dimensions. Nevertheless, adaptivity, i.e. the automatic run-time, or use-time adaptation, lying at the heart of these systems, can be characterized by (the interaction of) two main distinct high-level processes, or phases, namely *interaction assessment* and *adaptation decision making*, as shown in Figure 1 (Karagiannidis, 1998).

In the interaction assessment phase, the aim is to reach high-level conclusions concerning the aspects of learner-computer interaction that are considered significant for the particular educational application. For example, assessment may detect that the learner has not understood a particular concept, has problems with the interface, etc. Assessment is usually based on “low-level” information that is provided through a monitoring mechanism, including, for example, keystrokes, answers to quizzes, etc.

In principle, the assessment process can take into account, i.e. address the detection of, several aspects of learner-computer interaction, including, for example, the educational material being presented, the tasks being performed, the machine capabilities, the

network connection bandwidth, etc. Nevertheless, in most existing systems, the assessment process focuses entirely on long- or short-term learner’s characteristics and the result of the assessment is integrated into a *learner model*, which captures information concerning the learner characteristics that are considered significant for a particular educational application (Karagiannidis et al, 1998). Adaptive educational hypermedia/hypertext applications, for example, usually take into account the learner’s goals, knowledge, background, experience and preferences (Brusilovsky, 1998).

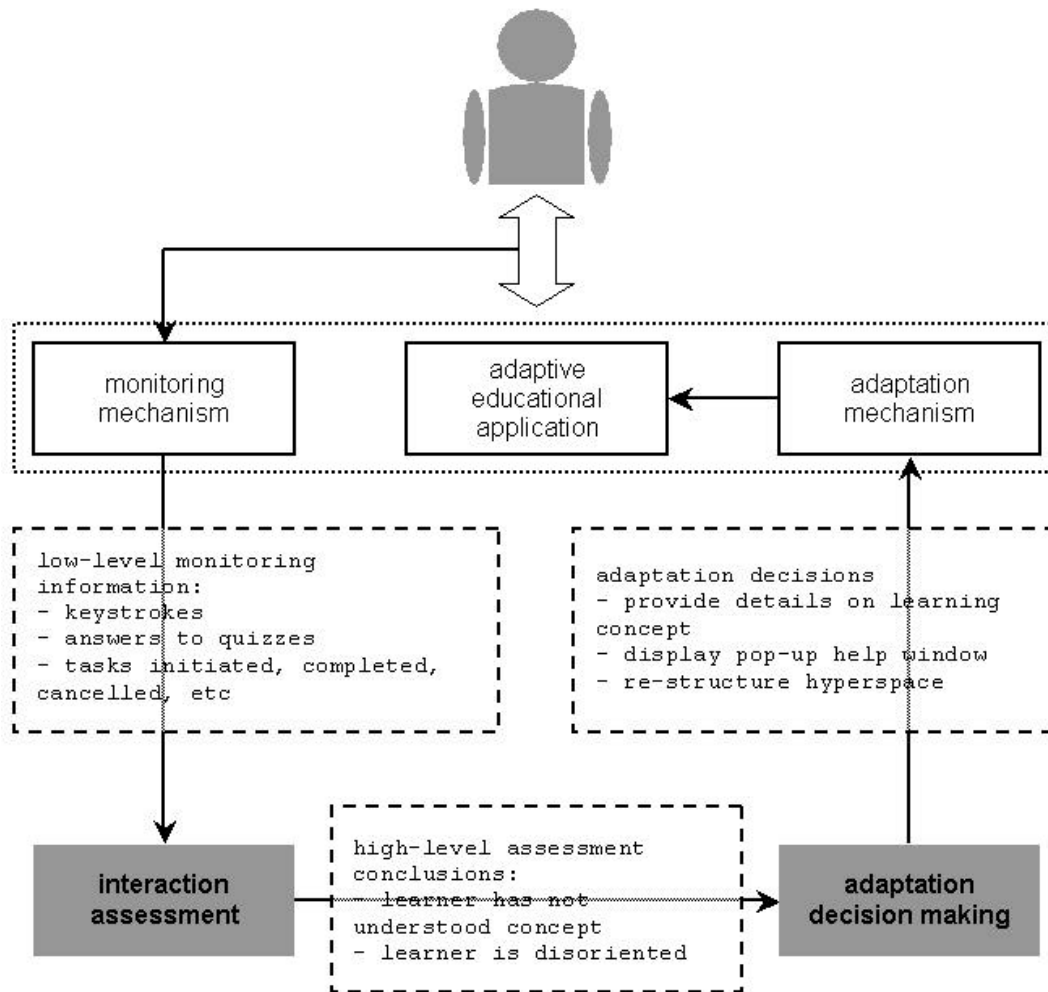


Figure 1 – Adaptation Decomposed.

We consider that adaptation is characterized by (the interaction of) two high-level processes: *interaction assessment* and *adaptation decision making*

In the adaptation decision making phase, on the other hand, specific adaptations are selected, based on the results of the assessment phase, in order to “improve” selected aspects of the system. Adaptation decisions may, for example, result in the presentation of a pop-up message helping the learner; the re-structuring of the hyperspace helping the learner navigate in it; or the provision of additional explanation for a specific concept of the educational domain; etc.

The adaptation decision making is usually realized through, i.e. the adaptation logic is captured into, a set of *adaptation rules*, which determine which adaptation constituent(s) should be selected, according to the results of the assessment process. For example, in adaptive educational hypermedia/hypertext applications, these rules are responsible for adaptive – text and/or multimedia – presentation, and/or adaptive navigation support, including the sorting, hiding and annotation of links (Brusilovsky, 1998). Implicit to these rules are the adaptation goals, and their possible trade-offs, which may vary considerably according to the requirements of the application.

The above processes are closely *interconnected*, since adaptation decision making strictly relies on - i.e. takes input from – the results of the interaction assessment. On the other hand, they are also *independent and orthogonal*, since the same assessment conclusions may result in significantly different adaptation decisions. (Karagiannidis et al, 1997a; Karagiannidis et al, 1997b).

The current practice in the evaluation of adaptive educational applications and services usually adopts a “with or without adaptation” approach, where experiments are conducted between two groups of learners, one working with the adaptive application, the other with its “non-adaptive version” – presuming, of course, that an adaptive educational application can be easily decomposed into its “adaptive” and “non-adaptive” components. In this sense, adaptive educational applications and services are usually evaluated “as a whole”, i.e. the evaluation process focuses on the overall learner’s performance, or the learner’s satisfaction, according to selected criteria. While this is reasonable, since the overall criterion for evaluating interactive systems is – or should be – the learner’s satisfaction, or the learner’s performance based on selected, measurable criteria, this approach does not provide useful information concerning the improvement of a system that is not found to be satisfactory. In particular, since adaptive behavior is evaluated as a whole, the reasons behind unsatisfactory adaptive behavior – when found – are not evident, and the necessary corrections are not clear. Moreover, evaluation results cannot be easily generalized, and successful design practices cannot be easily re-used across different applications and services (Karagiannidis & Sampson, 2000).

In this context, layered evaluation, where the different aspects that affect adaptation are evaluated separately, has been proposed as an effective means for evaluating adaptive applications and services (Brusilovsky et al, 2001). This paper advocates the employment of layered evaluation for adaptive and personalized educational applications and services. The paper then demonstrates the benefits of this approach through two specific examples: we outline how layered evaluation can improve the (forthcoming) evaluation of the KOD system; and we outline how the (already conducted) evaluation of the InterBook system can be revisited in the light of layered evaluation.

2 Layered Evaluation – The Framework

2.1 Layer 1 – Interaction Assessment Evaluation

In this layer, only the assessment process is being evaluated. That is, the question here can be stated as: “*are the conclusions drawn by the system concerning the characteristics of the learner-computer interaction valid?*”; or “are the learner’s characteristics being successfully detected by the system and stored in the learner model?”.

For instance, in the case of adaptive hypermedia systems, following the classification found in (Brusilovsky, 1998), this layer addresses the following issues: does the system detect the real learner goals, as they are continuously changing? is the learner’s actual knowledge of the system being successfully detected? are the learner’s interests actually detected by the system? is the learner’s experience with respect to the hyperspace structure successfully reflected in the learner model? are the learner’s preferences successfully represented in the learner model?

This phase can be evaluated, for example, through user tests, where experts can monitor learners as they work with the system, comparing their expert opinion on the learner’s characteristics versus the conclusions that are stored in the learner model. Additionally, learners can also themselves evaluate whether the conclusions drawn by the system at any particular instance reflect their real needs: “the system detected that my goal, at a particular instance, had been to know more about this subject; was this really the case?”. Moreover, this evaluation layer does *not* require that the adaptation decision making phase has been developed, i.e. the adaptive system has been fully developed.

Given that the assessment process has been evaluated separately and found satisfactory, its results can be generalized. That is, we can argue that the conclusions made by the assessment process, based on the low-level monitoring information, can be re-used in similar contexts, i.e. even with different decision making modules. This can facilitate the re-use of successful “design practices”, i.e. the logic underlying the interaction assessment process.

2.2 Layer 2 – Adaptation Decision Making Evaluation

In this case, only the adaptation decision making is being evaluated. That is, the question here can be stated as: “*are the adaptation decisions valid and meaningful, for selected assessment results?*”. Again, following the classification of (Brusilovsky, 1998), the above can be exemplified as: is the selected adaptive presentation technique appropriate for the given learner goals? does the selected adaptive navigation technique improve interaction, for specific learner’s interests, knowledge, etc?

This phase can, again, be evaluated through user testing, based on specific scenarios, where, for example, learners are given a particular goal, and it is evaluated whether the specific adaptation chosen helps with this goal. Again, learners and experts can evaluate whether specific adaptations contribute to the quality of interaction: “does the selected adaptation of the presentation of information improve the quality of the system, when the

learner is disoriented?”. Moreover, as in the previous case, this evaluation layer does *not* require that the interaction assessment phase has been developed, i.e. the adaptive system has been fully developed.

Again, given that the decision making phase has been evaluated separately and found successful, we can generalize its results. That is, we can argue that the design practice adopted in the particular application, as this is reflected in the adaptation logic - i.e. the adaptation rules – can be re-used across similar applications, even with different assessment processes.

3 Examples of Layered Evaluation

3.1 Layered Evaluation of the KOD System

The main objective of the KOD system is the delivery of an adaptive learning environment for personalized learning (see acknowledgements section). One of the major operational objectives of the project is to build on, and extend existing and emerging international *e*Learning standards, undertaken within the IEEE Learning Technology Scientific Committee (LTSC - ltsc.ieee.org), the IMS project (Instructional Management Systems project - www.imsproject.org), etc.

Existing *e*Learning standards and specifications already encompass most aspects of a standard *e*Learning architecture (e.g. the LTSA, Learning Technologies Standard Architecture), from the description of learning objects meta-data based on shareable XML based data structures (i.e. IEEE LOM, Learning Objects Metadata Schemas), to the assessment of learner performances (i.e. QTI, Question and Testing Interoperability Schemas), including standard multimedia components wrapping and delivery (i.e. Content Packaging and content API interfacing to underlying CMI, Computer Managed Instruction system). However, these specifications are still rather inadequate when it comes to support the definition and interchange of *reusable* adaptive and flexible learning methods, beyond the rigid approach of directive, curricular based, “linear” learning.

In particular, the current version of the Content Packaging Specification enables users, publishers, system managers, etc, to “package” and publish content structures built on content building blocks: the Content Packaging XML schema includes an “organization” field, describing the content structure included in the package. Currently, this field, although open to any notational description of navigation, mainly considers “rigid”, hierarchical, tree-based content structures description; no standard declarative notation, toolkit and viewer is addressed for conditional branching navigation, or for path redirection.

In this context, the KOD project is currently working on an extension of the Content Packaging Specification, so as to enable users and publishers to define, and share, not only content and content routes, but also navigation algorithms (i.e. conditional branching based on learner profiles). That is, to enable the definition of conditions in the

“organization” field of the Content Packaging Specification, which will define when specific learning content will be presented to learners, according to their profiles.

According to the above description, the main idea behind the KOD system is to “package” learning content in such a way that it can be “disaggregated” in a different way for different learners. The demonstration phase of the project involves the development of such a “KOD package” in the field of tele-medicine, which will also form the basis for the assessment and evaluation phase. As it is evident, this “packaging” of the learning material requires expertise in the learning content (i.e. tele-medicine in this case), as well as in instructional design, especially for the definition of the rules which will be imported in the KOD package for driving its conditional disaggregation for meeting the requirements of diverse learners.

Following the “traditional” evaluation methods of adaptive learning environments, the assessment of the KOD system would be conducted as follows. The KOD system would be installed in the demonstration site, and two sets of experiments would be performed:

1. one group of learners would work with the KOD (adaptive) system, which would access the “KOD tele-medicine package”; the KOD system would monitor each learner (i.e. use a learner profile), disaggregate the KOD tele-medicine package, parse the navigation rules that are included in it, and present to each learner only the learning content that is appropriate for his/her profile, according to the navigation rules.
2. the same, or a different group of learners would then access the “traditional” tele-medicine content (i.e. not packaged through KOD), through one of the existing (non-adaptive) *eLearning* platforms.

Both groups would be then assessed according to pre-selected criteria (e.g. answer to quizzes on tele-medicine), so as to evaluate whether, for example, learners of the 1st group (working with KOD) performed better than those not working with KOD.

If the assessment phase indicates that KOD is “superior” (e.g. in terms of learning effectiveness) of the non-adaptive *eLearning* platform, then the adaptation of the KOD system will be considered successful. If, however, the KOD system is found less effective, then we cannot be able to identify the “source” for this unsatisfactory result:

1. it could be the case that the learner model of the KOD system is not satisfactory; i.e. that the conclusions made by the KOD system for learners’ background, preferences, etc, are not correct;
2. it could also be the case that the KOD learner model is satisfactory, but the (instructional design) rules included in the content package are not successful

In contrast, adopting the layered evaluation approach would practically mean that (i) the success of the learner model (interaction assessment), and (ii) the navigation rules included in the KOD package (adaptation decision making), would be evaluated separately. As a result:

1. in case that the KOD system is found unsatisfactory, we will be able to identify which of the two components needs to be improved;
2. moreover, in case that the KOD system is found satisfactory, we can safely argue that both components are satisfactory; moreover, we could, in specific contexts, re-use these components; for example, we could re-use the navigational algorithms, which, in any case, are very expensive to define.

3.2 Layered Evaluation of the InterBook System

Recently we have tried to reconstruct an earlier study of adaptive annotation in the InterBook system using layered evaluation approach. We think that this study can clearly demonstrate the need and the benefits of layered evaluation. The study was originally reported in (Brusilovsky & Eklund, 1998). The goal of this experiment was to assess what impact, if any, user model-based link annotation would have on students' learning and on their paths through the learning space. Contrary to our expectations, the study *brought no significant results*. In particular, while students seem to understand and like adaptive navigation support (ANS) features, it didn't influence their performance on tests.

The question that is usually explored by the experimenters in such a situation is: *are there still any differences between adaptive and non-adaptive versions?* It is exactly the question we have tried to answer in the original report of the study (Brusilovsky & Eklund, 1998). However, from the prospect of a layered evaluation presented in this paper, different questions need to be considered such as: *Why does the adaptation not work? Was it the interaction assessment part where the system has performed poorly? Was it the adaptation decision making part where the adaptation decisions weren't properly made? Or, maybe the system was far from perfection in both layers of adaptation?* A layered evaluation approach could provide answers to these questions and guidance for further work.

In our case we were not planning a layered evaluation in advance, however we made a wise decision to collect lots of data about student interaction (more than we were expected to use). In this situation it became possible to perform a limited layered evaluation "post-factum" by re-processing the data. We have decided to check whether the educational status of a page (i.e. ready, not ready, or nothing new) predicted by the system has any connection with their performance on the page. The parameter we have checked is the average time spent by a user on pages of each of the three possible types (these data could be obtained by re-processing InterBook log files. It turned out that the average time students spent on "nothing new", "not ready", and "ready" pages are very different. The average time spent on a not-ready page is much larger than the time for a ready page, which is close to the average time per hit. The average time spent on a "nothing new" page is much less than average time per hit (Figure 2). This data shows that the interaction assessment process, which predicts an educational status of electronic pages, works quite well. A page classified as "nothing new" can be read much faster (or just passed over) because it has no new information, and a page classified as "not ready" is the most hard to understand because some background may be missed

While we have failed to make a correct conclusion when originally processing the data of our experiment, the work of other researchers provides some good evidence that this conclusion is, indeed, correct. An evaluation of the ELM-ART system (Weber & Specht, 1997), has shown that adaptive link annotation is of use for students who have some previous experience that is relevant to the subject being learned from an adaptive hypermedia system. In turn, novices benefit more from direct guidance with the adaptive “next” link. Similarly, Specht and Kobsa (1999) have shown that adaptive link annotation, a technology with little guidance and restriction, is a good way to help students with high previous knowledge on the subject (Specht & Kobsa, 1999). In turn, learners with low previous knowledge seem to profit from more guided and restrictive methods such as enabling/disabling links.

4 Discussion and Conclusions

This paper has outlined layered evaluation, and has argued that it can be an effective means for the evaluation of adaptive and personalized educational applications and services. It has also demonstrated the benefits of this approach through two specific examples: how the forthcoming evaluation of the KOD system can be improved by adopting this approach; and how a previous study of the InterBook system can be revisited in the light of layered evaluation.

We believe that the examples presented in this paper demonstrate the benefits of layered evaluation. Especially for the InterBook evaluation, which has already been performed, we really wish we were using the layered approach with full understanding during planning and performing our original study. It could have lead us to the correct conclusion right away. It could have pushed us to collect some more data about students (such as Web experience, level of education, etc) and possibly isolate a subgroup for which the selected method of adaptation may work. For this study, we could only reconstruct it and re-interpret its results with the layered approach at hand. We intend to use the layered approach for our future studies and we hope that the case described in this paper will convince other learner modeling researchers to also adopt this approach.

Acknowledgements

Part of the work presented in this paper was partially financially supported by the European Commission under the IST No 12503 Project “KOD – Knowledge on Demand” (<http://www.kodweb.org>, <http://kod.itl.gr>) through the Information Society Technologies Programme (IST).

Part of the ideas presented in this paper (in particular, the decomposition of adaptation into the interaction assessment and the adaptation decision making phases) have been initiated while the 2nd author was with the Assistive Technology and Human-Computer Interaction Laboratory of the Institute of Computer Science, Foundation for Research and Technology – Hellas (I.C.S.–FO.R.T.H.), Greece, from 1993 to 1998.

References

- (Brusilovsky, 1998) Brusilovsky P., “Methods and Techniques of Adaptive Hypermedia”, in Brusilovsky P., Kobsa A. & Vassileva J. (eds.), *Adaptive Hypertext and Hypermedia*, Kluwer Academic Publishers.
- (Brusilovsky & Eklund, 1998) Brusilovsky P. & Eklund J., “A study of user-model based link annotation in educational hypermedia”, *Journal of Universal Computer Science*, Special Issue on Assessment Issues for Educational Software, 4 (4).
- (Brusilovsky et al, 2001) Brusilovsky P., Karagiannidis C. & Sampson D., “A Case for Layered Evaluation of Adaptive Applications and Services”, In Review.
- (Karagiannidis et al, 1997a) Karagiannidis C., Koumpis A. & Stephanidis C., “Modeling Decisions in Intelligent User Interfaces”, *International Journal of Intelligent Systems*, 12 (10).
- (Karagiannidis et al, 1997b) Karagiannidis C., Koumpis A. & Stephanidis C., “Adaptation in Intelligent Multimedia Presentation Systems as a Decision Making Process”, *Computer Standards and Interfaces*, Special Issue “Towards a Standard Reference Model for Intelligent Multimedia Presentation Systems”, 18(6-7).
- (Karagiannidis, 1998) Karagiannidis C., *Supporting Run-Time Adaptation in Intelligent User Interfaces: Assessment of Interaction and Design of Adaptation*, Ph.D. Thesis, University of Kent at Canterbury, U.K.
- (Karagiannidis et al, 1998) Karagiannidis C., Koumpis A., Stephanidis C. & Georgiou A.C., “Employing Queuing Modeling in Intelligent Multimedia User Interfaces”, *International Journal of Human-Computer Interaction*, 10 (4).
- (Karagiannidis & Sampson, 2000) Karagiannidis C. & Sampson D., “Layered Evaluation of Adaptive Applications and Services”, In Brusilovsky P., Stock O. & Strapparava C. (eds), *Adaptive Hypermedia and Adaptive Web-Based Systems*, Proceedings of the AH2000 International Conference, Trento, Italy, August 2000, Lecture Notes in Computer Science No 1892, Springer Verlag.
- (Specht & Kobsa, 1999) Specht M. & Kobsa A., “Interaction of domain expertise and interface design in adaptive educational hypermedia”, *2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web*.
<http://wwwis.win.tue.nl/asum99/specht/specht.html>
- (Weber & Specht, 1997) Weber G. & Specht M., “User modeling and adaptive navigation support in WWW-based tutoring systems”, *6th International Conference on User Modeling* (pp. 289-300).