Adaptive E-learning Systems: Evaluation Issues

Cristina Hava Muntean, Jennifer McManis

School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland
Fax: +353 1 7005508, E-mail: havac@eeng.dcu.ie, mcmanisj@eeng.dcu.ie

Abstract *- The large variety of user-adaptive educational hypermedia systems available requires techniques for evaluation and comparison with other adaptive systems. In this paper the current evaluation approaches used for assessing intelligent e-learning environments are presented. The paper discusses important issues emerging from research on system evaluation that focuses on both social and practical acceptability of courseware applications. Acceptability of an educational system is analysed in terms of usability evaluation, learner achievement and learning performance evaluation.*

Keywords: *Evaluation of adaptive e-learning systems, usability, Web-based adaptive hypermedia systems*

I. INTRODUCTION

User-adaptive systems are interactive software systems that automatically adapt to properties and behaviors of individual users. Research in this area has also associated the term user-adaptive system with adaptive user interfaces, user modeling, and personalization. The educational domain was the first application area to see the introduction of different user-driven adaptive techniques before the research extended to other areas such as tourism, e-commerce, health care, etc. With the increase in the popularity of the Internet, the Web started to have an important influence on teaching and learning styles today, mainly in higher education. Therefore, many online lecture notes or complex tutoring applications were distributed on the Web.

In order to increase the usability of the hypermedia (hyperspace) and to bring flexibility and personalization capability of the conventional presentation to the materials on the Web, a new research direction within the area of user-adaptive systems has appeared: Adaptive Hypermedia. In this context Adaptive Hypermedia Systems (AHS) build a model of the goals, preferences and knowledge of the individual user and use it in order to perform adaptations to the needs of the user. In general, an AHS consists of a user model, which describes information about the user, a domain model that describes the domain subject and an adaptation model, which describes how the adaptation should be performed. Users' needs are met through two types of adaptation: adaptive navigation support and adaptive presentation [1]. AHS are used now in several application areas where the hyperspace is reasonable large and is expected to be used by individuals with different goals, knowledge and backgrounds [2]. The state of the art in adaptive hypermedia and more details on adaptive methods and techniques used by these systems are presented in [1, 3].

Adaptive Hypermedia Systems for Education (AHSE) in general and mainly web-based AHSE, have attracted considerable interest due to their huge potential to facilitate personalized learning and in consequence many intelligent e-learning systems were proposed. Among them AHA! [4, 5], ELM-ART [6], InterBook [7] and ISIS-Tutor [8] are the most well known.

However in the adaptive hypermedia area there is a significant lack of assessment and evaluation strategies, comprehensive empirical studies to measure both the usefulness of adaptation within the systems and between the systems, and systems' decision-making capabilities. There is also much debate on how adaptive hypermedia applications should be evaluated since there is no standard or agreed evaluation framework for measuring the value and the effectiveness of adaptation yielded by adaptive systems.

The objective of this paper is to address the assessment strategies and evaluation methods used in web based learning environments and adaptive hypermedia. The paper presents a survey of the research that has been undertaken in order to evaluate adaptive educational applications with emphasis on we-based AHSE.

II. CURRENT EVALUATION APPROACHES

The current, and most used method in the evaluation of adaptive educational systems adopts a **"with or without adaptation approach"** [9], considering that the evaluated system can have adaptive and non-adaptive versions. The experiments are conducted between two groups of learners, one working with an adaptive version of the system and the other with its non-adaptive version. This conventional method of comparing an adaptive and non-adaptive version

of an application is debatable [10] and highly depends on how the non-adaptive version was obtained.

A possibility is to "disable" all adaptive features of the adaptive version [10]. Since most adaptive systems are developed with particular adaptive techniques in mind, removing those techniques affects the system's basic functionality and the comparisons are always in favour of the adaptive systems. Although highly used this approach does not offer fair results.

Another possibility is for this comparison to be performed with the original non-adaptive system prior to adding adaptive functionality. This lacks the advantage of a well-structured domain model as in the adaptive version and may lead again unfair results according to some opinions. This is especially since both the information content of the pages and the link structure and/or presentation layout may be different in the two versions. Short explanations, additional details comparisons can be added, changes in the presentation style can be performed and/or presentation length can be modified. Links or link destinations can be added, removed, sorted, annotated [10].

A third possibility is to disable some adaptive features from the adaptive version of the system. This allows for the comparison to be made between two adaptive versions of the same system, having different degree of adaptiveness. This type of comparison is used to show the benefits of some adaptive techniques against others.

From the evaluation strategy point of view, two main directions were proposed.

A. System Evaluation "As a Whole"

The first approach targeted **adaptive system evaluation "as a whole"** [11] and is very often used in educational area. The evaluation process focuses mainly on the overall learners' performance and their satisfaction related to the use of the adaptive system. This user satisfaction can be quantified by selected and measurable criteria. In this context the most used criteria in the evaluation process of educational systems are: task completion time, learning performance assessed by comparing the results of a pre-test and post-test, number of navigation steps, number of times the subjects revisited "concepts" they were attempting to learn, users satisfaction reflected through questionnaires [7, 8, 12, 13].

B. Layered Evaluation of the System

Very recent, a new approach was recommended for the evaluation of the adaptive applications and advocated by a number of researchers [11, 14, 15, 16, 17]. This approach is based on **layered evaluation of adaptive applications**. Unlike in the previous approach that focuses on assessing user's performance and satisfaction in relation with the system as a whole, layered evaluation assesses the success of the adaptation by decomposing the system into different layers and evaluating them one by one [18]. The different layers reflect various aspects and stages of the adaptation. Although the current proposed frameworks are described at different levels of granularity [15, 19], mainly the evaluation process is divided in two phases: evaluation of the interaction assessment phase and evaluation of the adaptation decision-making phase [14].

Karagiannidis et. al [14] has proposed a framework for layered evaluation that consists of two layers:

- *Layer 1: Interaction Assessment Evaluation* that tests if the system detected the learner's goals, knowledge, preferences, interests, user's experience with the respect of hyperspace. It also assesses whether the assumption drawn by the system concerning characteristics of the user-computer interaction is valid. This evaluation is based on comparison between experts' opinions and information stored in learner (user) model.
- *Layer 2: Adaptation Making Evaluation* that tests if the selected adaptive technique is appropriate, valid and meaningful for learner's goal or improves interaction for specific learner's interests, knowledge, etc. This evaluation consists of tests based on scenarios that involve a particular goal for the learner and assess the success of quality improvement. Learners and/or experts can evaluate the tests.

The division of the evaluation process into the two layers that also reflect the main phases of the adaptation may help to determine where the fault (if any) of the adaptive system may be and to target the solutions accordingly [11]. For example it can be the case that adaptation decisions are reasonable but they are based on incorrect system assumptions, or that the system assumptions are correct but the adaptation decision is not meaningful. Both cases can also happen at the same time.

A more detailed approach was proposed by Weibelzahl et. al [15] and consists of a framework for layered evaluation based on four layers:

- *Layer 1: Evaluation of the Reliability and Input Data*. This evaluation prevents unreliable input data to result in miss-adaptation.
- *Layer 2: Evaluation of Inference*. This layer evaluation test the inference mechanism in different environments under real world conditions
- *Layer 3: Evaluation of Adaptation Decision*. The idea of the evaluation is that if some user properties have been inferred, several adaptation possibilities exist. (e.g. with/without adaptive guiding, with/without link annotations).

- *Layer 4: Evaluation of Interaction*. In this case human system interaction has to be evaluated to prevent confusion and dissatisfaction of the users. Different objective and subjective measures are taken into account such as: system usability, solution quality, frequency of tasks success, number of required hints, etc.

One can notice that both evaluation strategies: evaluation "as a whole" and layered evaluation aim at assessing three important features of the educational applications: usability of the application, learner achievement and learning performance.

In the following section methodologies of assessing these features related to intelligent e-learning systems are presented in more details.

III. EVALUATION TESTS OF ADAPTIVE EDUCATIONAL SYSTEMS

A. Usability Evaluation Tests

One of the most important features of any software application is its usability. According to ISO 9241 standard, usability represents the effectiveness, efficiency and satisfaction that a software application offers to its users in a given context of use and task.

In an educational environment the usability of software application is related to its pedagogical value [20]. Although there is a large amount of knowledge relating to educational software usability evaluation strategies [21], currently there are not well-defined techniques for usability evaluation of e-learning (distance learning) environments [22]. This is due to the fact that e-learning is an area of relatively short history, users of e-learning tools can access them through various computer, network and social contexts and the characteristics of a typical users of e-learning services can not be easily predicted [23].

Some of the most used methods proposed in the literature to be applied during the usability evaluation are: query techniques (interviews and questionnaires), logging of user performance in laboratory conditions, timing and keystroke level measurements, subjects' observation through adequate equipment, heuristic evaluation, etc. These methods are applied after the subjects have interacted with the system by performing one or multiple tasks. Usually the usability is analysed through five major characteristics: efficient to use, easy to remember, pleasant to use, easy to learn, few errors.

Questionnaires and interviews are the most widely used technique since they provide a quantitative measure of usability and they serve as an objective comparison of two systems. This technique offers a concise test of usability, it gets directly the users' viewpoint and attitude and it is suitable for wide range of end-users, especially students. A big advantage is that it does not require the presence of an evaluator. In this context, Preece [24] suggested a list of guidelines for creating questions for the questionnaires, currently widely used for the usability evaluation of the web-based systems.

Heuristic evaluation is also a widely accepted method for diagnosing the system's usability due to the fact that it can be completed in a relatively short period of time. This methodology involves an expert that evaluates the system using a set of recognized usability principles, called "heuristics" by Nielsen [25].

B. Learner Achievement Evaluation

In the evaluation of learning process quality and quantity of learning (learning outcome) is very important to be assessed. Therefore learner achievement (defined as the degree of knowledge accumulation by a person after studying a certain material) continues to be a widely used barometer for determining the utility and value of learning technologies. It is analysed in the form of *course grades, pre/post-test scores, or standardized test scores*.

A course grade is a certification of competence that should reflect, as accurately as possible, a student's performance in a course. There are multiple methods for assigning grades, such as weighting, distribution gap method, curve, percent grading, relative grading, and absolute standard grading.

Pre/Post test scores are also a viable methodology to assess the extent to which an educational intervention has had an impact on student "learning". Pre-test is used to determine subject's prior knowledge on the studied domain, while post-test is used to examine learning outcomes after the intervention.

Standardized tests scores give a "standard'" of measure of students' performance when a large numbers of students often geographically distributed take the same test.

Tests, quizzes or exams are methods used to evaluate students and assess whether they learned what it is expected to be learned. Jacobs and Chase [26] made a distinction between the three terms: tests, quizzes and exams, based on the scope of content covered. An examination is the most comprehensive form of testing. A test is more limited in scope, focusing on particular aspects of the course material. A quiz is even more limited and usually is administered in fifteen minutes or less.

Among them, tests are the most important one for the evaluation of adaptive Web-based learning systems, for two main reasons [6]:

1) Testing offers a feedback on the correctness of the answers, helping to optimise learning process.

2) Testing results are the most reliable source of evidence that a user has learned a concept.

Tests- quizzes- or exams -based evaluation may consist of five different types of test items:

1) Yes-No (True-False) test items: users have to answer to questions by selecting Yes or No answer only.
2) Forced-Choice test items: users have to answer a question by selecting only one of the alternative answers.
3) Multi-Choice test items: users have to answer a question by selecting all correct answers provided.
4) Essay (Free-Form) /Short Answer test items: users can type an answer to the question asked freely into the form. Short answers are usually only one to three paragraphs long.
5) Gap-Filling (Completion) test items: users have to type in characters or numbers to complete a word or sentence.

Each type of test items has its relative strengths and weaknesses and they are discussed next.

Yes-No (True-False) tests: Measure the ability to identify the correctness of statements of facts, definitions of terms, statements of principles, etc. These tests can sample many more bits of information in a given time period than any other type of test format. Research does indicate true-false testing is sufficiently reliable and valid for periodic in-classroom testing. Because of random guessing (50-50 chance), these tests can be less reliable than other tests, unless the number of questions asked is high.

Forced-Choice tests and Multi-Choice tests: These types of tests consist of a stem that describes a problem and a series of possible answers or alternatives (usually 3 to 5). They can address many learning targets, can be used to assess both simple knowledge and complex concepts and can be answered quickly. They are easy to score, can be considered objective because all potential item responses are identified, but lack the ability to address learner produced responses. Multi-choice tests have a higher degree of difficulty than the forced-choice tests and both are more difficult then Yes/No tests.

Essay/Short Answers tests: These tests are most advantageous when assessing complex learning outcomes and higher-level thinking skills. They are relatively easy to construct, do not permit guessing and cannot be answered by simply recognizing the correct response. Among the limitations of Essay Tests are that they are difficult to score, their scores are less reliable than well written objective tests, the score is influenced by the readers overall impression of the student and they provide a very limited sample of the content in the typical unit of study.

Gap-Filling (Completion) tests: They make scoring faster and less subjective. They are used to measure the recall of memorized information. Completion test items preclude the kind of guessing that is possible on limited-choice items since they require recall and a definite response rather than simple recognition of the correct answer. They are more difficult to score than forced-choice items and scoring often must be done by the test writer since more than one answer may be considered correct. On the whole, completion items have little advantage over other item types unless the need for specific recall is essential.

Every type of test has a general value for difficulty and relevance for the tested concept. These test items can be used to wrap up a course, lesson, section, or subsection.

C. Learning Performance Evaluation

Learning performance term refers to how fast a study task (e.g. learning task, searching for a piece of information or memorising information displayed on the computer screen) takes place. The most used metric used for measuring the learning performance provided by an adaptive hypermedia system for education is *study session time* [6, 12, 27, 28]. The completion time for a study session is measured from the start of the session, when the subject logins into the system and starts to study, until the subject starts answering the questions from the evaluation test. Other metrics worth to be mentioned are: *number of navigation steps* performed during a study session [6, 8, 12, 13], *number of pages re-visited* [8], *average time spent per page* for studying the information, *average access time*.

D. Assessment of the Evaluation Results

The assessment of the usability evaluation is performed in terms of overall usability of the web-based course system and usability of each category of questions that reflects different characteristics of the system such as efficient to use, easy to remember, pleasant to use, easy to learn, few errors, etc. Mean values and standard deviations of the results are computed.

The assessment of the learner achievement is performed in terms of final scores from the quizzes, tests or exams, achieved by the subjects when one or more versions of the adaptive educational system is used. The results are analysed by computing mean values and standard distributions based on the final scores

Learning performance is analysed through the measured performance metrics (e.g. study session time).

For scientific credibility, different statistical methods for data analysis are used for the comparison of the two or more versions of adaptive systems. The most used statistical analysis methods in the evaluation of educational systems or hypermedia systems [12, 23] are:

T-Tests

It is the most widely used statistical test of all time because it is simple, straightforward, easy to use, and adaptable to a broad range of situations. The t-test allows analysing if there is a statistically significant difference between the means of two groups, at a certain confidence level. In HCI practice t-test is also commonly used to compare how groups of subjects perform in two different test conditions. T-Test analysis involves: the definition of the null hypothesis and of the significance level of the test (typically stated at the 0.05 – 0.01 level), and the computation of t-value and df-value (degrees of freedom). The null hypothesis of a t-test always states that the results of the two groups do not differ significatively. The t-test is used to prove or to discard the mentioned null hypothesis.

ANOVA Tests

The previous paragraph has shown how T-Test is used to compare means from two independent groups. ANOVA (ANalysis Of VAriance) Test is used to compare means from *k* independent groups, where *k* is 2 or greater. In fact, T-Test is considered to be a special two-group version of ANOVA. The null hypothesis of ANOVA-Test states that means from two or more samples are equal, while the alternative hypothesis states that at least one population mean differs. By performing the test one of the hypotheses is rejected and the other one accepted.

F-Test

It offers a statistical analysis of the equality of the two population variances related to precision and accuracy. It allows deciding if the two variances are comparable with a certain confidence level. F-Test analysis involves: the definition of the null hypothesis and of the significance level and the computation of variance (SD^2) and f-value. The null hypothesis of a F-test always states that the there are not statistical difference in precision/accuracy. The t-value is used to prove or to discard the mentioned null hypothesis.

Q-Test

This test is very useful when in the data set there are one or more values, which appear to be anomalous. Therefore, Q-test allows determining if a very low or very high measurement can be discarded. Q-Test analysis involves: the definition of the null hypothesis, of the anomaly value, the number of the total observations performed, of the significance level, and the computation of q-value. The null hypothesis of a G-test always states that the measurement is statistically important and cannot be rejected. The q-value is used to prove or to discard the mentioned null hypothesis with the given confidence level.

IV. CONCLUSIONS

The goal of this paper was to present the current assessment strategies and evaluation methods used in adaptive e-learning environments. The evaluation methodologies aim at assessing the most important characteristics: usability, learner achievement and learning performance that validate the acceptability of an educational system.

These assessment strategies were used for the evaluation of QoSAHA, a performance oriented learning system. More details about the QoSAHA system and preliminary results of the evaluation tests were presented in [29, 30].

V. ACKNOWLEDGEMENTS

The support of the Informatics Research initiative of Enterprise Ireland is gratefully acknowledged.

REFERENCES

[1] P. Brusilovsky, “Adaptive hypermedia”, User modeling and user adapted interaction”, Vol. 11, pp. 87-110, 2001.

[2] P. Brusilovsky, “Methods and techniques of adaptive hypermedia”, Journal of User Modeling and User-Adapted Interaction, Special Issue on adaptive hypertext and hypermedia, Vol. 6, No. 2-3, pp. 87-129, 1996.

[3] A. Kobsa, J. Koenemann and W. Pohl, “Personalised hypermedia presentation techniques for improving online customer relationships”, The Knowledge Engineering Review, Vol. 16 No. 2, Cambridge University Press, pp. 111–155, 2001.

[4] P. De Bra and L. Calvi, “AHA! An open adaptive hypermedia architecture”, Journal of The New Review of Hypermedia and Multimedia, Vol. 4, No. 1, pp.15-139, 1998.

[5] P. De Bra, “Adaptive educational hypermedia on the Web”, Journal of Communications of the ACM, Vol. 45, No. 5, pp. 60-61, 2002.

[6] G. Weber and P. Brusilovsky, “ELM-ART: An adaptive versatile system for Web-based instruction”, International Journal of Artificial Intelligence in Education, No. 12, pp. 351-384, 2001.

[7] P. Brusilovsky, J. Eklund and E. Schwarz, “Web-based education for all: A tool for developing adaptive courseware”, Journal of Computer Networks and ISDN Systems, Vol. 30, No. 1–7, pp. 291–300, 1998.

[8] P. Brusilovsky, “Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-Tutor”, Journal of Computing and Information Technology, Vol. 6 No. 1, pp. 27-38, 1998.

[9] C. Karagiannidis, D. Sampson and P. Brusilovsky “Layered evaluation of adaptive and personalized educational applications and services”, 10th International Conference on Artificial Intelligence in Education, Workshop on Assessment Methods in Web-based Learning Environments and Adaptive Hypermedia, San Antonio, Texas, US, 2001.

[10] P. De Bra, “Pros and cons of adaptive hypermedia in Web-based education”, Journal on Cyber Psychology and Behaviour, Vol. 3, No. 1, pp. 71-77, 2000.

[11] P. Brusilovsky, C. Karagiannidis, and D. Sampson, “The benefits of layered evaluation of adaptive applications and services’”, 18th

International Conference on User Modeling, Workshop on Empirical Evaluations of Adaptive Systems, Sonthofen, Germany, 2001

[12] M. H. Ng, W. Hall, P. Maier and R. Armstrong, "The application and evaluation of adaptive hypermedia techniques in Web-based medical education", Association for Learning Technology Journal Vol. 10, No. 3, pp. 19-40, 2002.

[13] C. Boyle and A. Encarnacion, "Metadoc: An adaptive hypertext reading system", User Modeling and User-Adapted Interaction Vol. 4, Kluwer Academic Publishers, pp. 1-19, 1994.

[14] C. Karagiannidis, and D. Sampson, "Layered evaluation of adaptive applications and services", International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2000), Trento, Italy, Springer LNCS 1892, pp. 343-346, 2000.

[15] S. Weibelzahl and G. Weber, "Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems", Künstliche Intelligenz Journal, Vol. 3, pp. 17-20, 2002.

[16] S. Weibelzahl, "Evaluation of adaptive systems", In Ma. Bauer, J. P. Gmytrasiewicz, and J. Vassileva, Eds, Proc. of the 8th International Conference on User Modeling (UM2001), Berlin, pp .292–294, 2001

[17] G. Elissavet and A. A. Economides, "An Evaluation instrument for hypermedia courseware", Journal of International Forum of Educational Technology & Society and IEEE Learning Technology Task Force, Vol. 6, No. 2, pp. 31-44, 2003.

[18] G. D. Magoulas, S. Y. Chen and K. A. Papanikolaou, "Integrating layered and heuristic evaluation for adaptive learning environments", 9th International Conference on User Modeling, Second Workshop on Empirical Evaluation of Adaptive Systems, Pittsburg, US, pp. 5- 14, 2003.

[19] E. Herder, "Utility-based evaluation of adaptive systems", 9th International Conference on User Modeling (UM2003), Second Workshop on Empirical Evaluation of Adaptive Systems, Pittsburg, US, pp. 25- 30, 2003

[20] D.Kirkpatrick**,** "Evaluating training programs", San Francisco, CA: Berrett- Koehler Publishers, Inc. 1994.

[21] N. M. Avouris, N. Tselios, & E.C. Tatakis, "Development and evaluation of a computer-based laboratory teaching tool", Journal of Computer Applications in Engineering Education, Vol. 9, 2001

[22] J. M. Heines, "Evaluating the effect of a course Web site on student performance", Journal of Computing in Higher Education, Vol.12, No. 1, pp. 57-83, 2000.

[23] N. Tselios, N. Avouris A. Dimitracopoulou, S. Daskalaki, "Evaluation of distance-learning environments: Impact of usability on student performance", AACE International Journal of Educational Telecommunications, Vol. 7, No. 4, pp. 355-378, 2001.

[24] J. Preece, "Online Communities: Designing usability, supporting sociability", published by John Willey & Sons, Chichester UK, 2000.

[25] J. Nielsen, "Heuristic evaluation. Usability inspection methods", Wiley, New York, 1994.

[26] L. C Jacobs and C. I. Chase, "Developing and using tests effectively: A guide for faculty", San Francisco: Jossey-Bass, 1992.

[27] A. Mitrovic, "Self-assessment: how good are students at it?", 10th International Conference on Artificial Intelligence in Education (AI-ED 2001), Workshop on Assessment Methods in Web-Based Learning Environments & Adaptive Hypermedia, San Antonio, Texas, US, 2001.

[28] N. Bajraktarevic, W. Hall and P. Fullick, "Incorporating learning styles in hypermedia environment: Empirical evaluation", ACM Hypertext Conference, Adaptive Hypermedia Workshop, Nottingham, UK, 2003.

[29] C. Hava Muntean, J. McManis, "QoSAHA: A performance oriented learning system", AACE ED-MEDIA 2004 Conference, Lugano, Switzerland, June 2004.

[30] C. Hava Muntean, J. McManis, "A QoS – aware adaptive Web-based system", IEEE International Conference on Communications (ICC 2004), Paris, France, June 2004.