USABILITY TESTING OF EDUCATIONAL SOFTWARE: METHODS, TECHNIQUES AND EVALUATORS

Ana Amélia Amorim Carvalho
University of Minho

aac@iep.uminho.pt

USABILITY TESTING OF EDUCATIONAL SOFTWARE:
METHODS, TECHNIQUES AND EVALUATORS

Ana Amélia Amorim Carvalho
University of Minho
aac@iep.uminho.pt

Key words: usability testing, methods, techniques, evaluators

Abstract - This paper stresses the importance of usability tests in user acceptance, performance, and satisfaction. We refer to methods, techniques, and evaluators used in usability testing. Then, we focus on the importance of planning a test and conducting it, attending to ethical issues. Finally, we describe usability tests conducted to an hyperdocument.

Introduction

One of the factors that affect the acceptability of educational software is its usability. Smith & Mayes (1996: 6) state that "usability is now recognised as a vital determining factor in the success of any new computer system or computer-based service". They go further and advise: "Manufacturers who have made insufficient efforts to build usability into simple technology such as video recorders and microwaves, have suffered in the market place. Consumers choose alternative, more effective products" (Smith & Mayes, 1996: 6).

Educational researchers (even Master or PhD students) should not overlook usability testing, if they want to develop educational software that is efficient, effective and gives satisfaction to the user. For achieving such specific aims it is worthy to know about usability methods, techniques, evaluators, when to apply usability tests, how to plan and conduct a test.

Usability: the concept

Human-Computer-Interaction (HCI) is the area where usability emerged. Several books or papers about HCI present a definition or characterization of usability. For instance, Hix & Hartson (1993) consider that usability is related to the interface efficacy and efficiency and to user reaction to the interface. Nielsen (1993[1], 1995) integrates usability as one of the parameters associated with system acceptability[2]. He associates five attributes to usability: easy to learn, efficient to use, easy to remember,

[1] Nielsen (1993) uses the following categories: learnability, efficiency, memorability, error, and satisfaction.

[2] The acceptability of a computer system is a combination of its social acceptability and its practical acceptability. If the system is socially acceptable, it is necessary to analyze its practical acceptability within categories such as cost, compatibility with existing systems, reliability, etc., as well as the category of usefulness. Usefulness is the issue of whether the system can be used to achieve some desired goal. It can be divided in two categories of utility (whether the functionality of the system can do what is needed or in an educational hypermedia students learn from using it) and usability (how well users can use that functionality).

few errors (the prevention of catastrophic errors is relevant for applications such as process control or medical applications), and pleasant to use.

Shackel (1990) refers to four aspects of interest in usability testing: learnability (easy of learn), throughout, flexibility, and attitude. Rubin (1994) accepts that usability includes one or more of the four factors outlined by Booth (1989): usefulness, effectiveness (ease of use), learnability, and attitude (likebility). For Smith and Mayes (1996) usability focuses on three aspects: easy to learn, easy to use and user satisfaction in using the system.

In international standards, usability refers to effectiveness and efficiency to achieve specified goals and users satisfaction. "Usability: the extent to which a product can be used by specified users to achieve a specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO/DIS 9241-11; European Usability Support Centres).

Based on these opinions about usability we may conclude that there are two broad areas to collect relevant data: system and user performance (efficacy, efficiency, easiness to learn and easiness to use) and user satisfaction in using it. In the remainder of this paper we focus on methods, techniques and evaluators for conducting usability tests. Then, we describe the usability tests carried out during the development of an hyperdocument about the Portuguese novel "Cousin Basilio" from Eça de Queirós, structured according to Cognitive Flexibility Theory.

Methods

Based on Preece (1993), we consider four usability evaluation methods: expert evaluation, observational evaluation, survey evaluation and experimental evaluation. The different methods imply different types of evaluators, different number of users, and different types of data to be collected.

Expert evaluation, also known as heuristic evaluation, is normally carried out by experienced people in interface design and human factors research who are asked to describe the potential problems they foresee for less experienced users. These experts often suggest solutions for the problems they identify. This method is efficient and provides prescriptive feedback. However, experts should not have been involved with previous versions of the prototype under evaluation and they should have suitable experience. The role of the experts needs to be clearly defined to ensure that they adopt the proper perspective when using the prototype. The tasks undertaken and the materials given to the experts should be representative of those intended for the eventual users. Finally, the form of reporting adopted by the expert needs to be specified so that information is obtained about the most important problems.

Observational evaluation implies collecting data that provide information about what users do when interacting with educational software. Several data collection techniques may be used. According to Preece (1993) two broad categories of data may be obtained: how users tackled the tasks given, where the major difficulties lie and what can be done; and performance measures like frequency of correct task completion, task timing, frequency of participant errors.

Surveys are employed to know users' opinions or to understand their preferences about an existing or potential product through the use of interviews or questionnaires.

In an *experimental evaluation* an evaluator can manipulate a number of factors associated with the interface and study their effect on user performance. It is necessary to plan everything very carefully: required level of user experience, hypotheses to be tested, the structure of tasks, time needed to complete the experiment, and so forth.

Other methods can also be applied such as: focus group, walk-through, paper-and pencil evaluations, usability audit, field studies, and follow-up studies (Rubin, 1994). Not all methods have to be used (Booth, 1989). Normally, a combination of methods is selected according to the needs and constraints of a project. The selection of a method has to take into account the techniques for data collection that can be applied.

Techniques

There are several techniques such as direct observation, video recording, software logging, interactive observation, verbal protocols, interviews and questionnaires associated with diverse methods (fig. 1).

Method	**Techniques**
Expert/ heuristc	Walk-through Questionnaires
Observation	Direct observation Video recording Software logging Verbal protocols (think aloud)
Survey	Interviews Questionnaires
Experimental	Software logging Questionnaires Interviews

Figure 1 - Methods and techniques to collect data

Direct observation involves observing users during task execution, with the evaluator taking notes about user performance and, if appropriate, timing sequences of actions. However, this technique can affect users' performance, as they may react as they think the evaluator/monitor would expect. This situation is known as Hawthorne effect. For avoiding such situation, *video recording* can be used. The recording is then replayed and users' behavior and problems are analyzed. Sometimes users and evaluators work together on the interpretation of protocols. Such participated evaluation is useful because users may explain some reactions or choices recorded. The participant may be invited to think-aloud (verbal protocol), when interacting with the software. By verbalizing their thoughts, users enable the monitor to understand how they view the product, and this makes it easy to identify the users' major misconceptions. From this technique a wide range of information can be obtained. However, users often find difficult to express their thoughts while trying to solve a difficult problem. Sometimes pairs of participants are used to avoid this situation, known as constructive interaction (Nielsen, 1993). In pairs, the test situation is much more natural, since they are used to verbalize when trying to solve a problem together.

Another useful technique is *software logging*. This technique records the interaction between the user and the educational software. The data is collected unobtrusively without influencing the user's working style. It usually consists of a time-stamped log of user input and software responses. It is possible to reconstruct exactly what the user was doing and the time spent on each feature, and also to analyze the frequency of use of certain features.

The *interview* is one way of collecting data in a survey. Interviews can be structured (sequence of predetermined questions with no exploration of individual attitudes) or flexible (it has some topics and develops in response to the interviewees' replies). The other way to collect data in a survey is through *questionnaires*. There are two types of questions: open questions (the respondent provides his/her own answer) and closed questions (the respondent selects the answer from a choice of alternative replies).

On the next section we focus on the third element of this evaluation triad: the evaluators.

Evaluators

Most of the researchers agree that there are two types of evaluators: experts and users (Nielsen, 1993; 1995; Preece, 1993; Rubin, 1994). Expert evaluations involve a review of an educational software, according to accepted usability principles. Nielsen (1992) indicated a "double" specialist, that is, an expert in usability who is also an expert in the particular technology employed by the software.

The observational, survey and experimental methods imply the presence of users. Participants have to be representative of the target population to evaluate the degree to which a product meets specific usability criteria.

An important issue for usability is users' individual characteristics and differences. There are three main dimensions along which users' experience is distinguished: experience with the system (from novice user of system to expert user of system), with computers in general (from minimal computer experience to extensive computer experience), and with the task domain (from ignorant about domain to knowledgeable about domain). These three dimensions are described as the "user cube" (Nielsen, 1993). Other aspects are important when characterizing the users such as age, gender, preferred learning styles, and other attributes that can be relevant for the particular educational software.

Tests should also address user performance and satisfaction. *Performance* data correspond to measures of participant behavior, focusing on aspects such as "efficiency and efficacy of use". It may measure: error rates, time to perform a task, number and percentage of tasks completed incorrectly, time spent reading a specific section, count of incorrect menu choices, count of incorrect icons selected, count of visits to the index, count of visits to the table of contents, count of negative comments, and so on. Nielsen (1993: 192) refers to "a major pitfall with respect to measurement is the potential for measuring something that is poorly related to the property one is really interested in assessing".

User Satisfaction also mentioned as preference data represent measures of participant opinion or thought process. It includes participant rankings, answers to questions, and so forth.

Rubin (1994) points out some aspects to measure, for example, usefulness of the product, how well product matched expectations, ease of use overall, ease of learning overall, ease of set up and installation, ease of accessibility, usefulness of the index, table of contents, help, graphics, and so on. User satisfaction can also be measured through a comparison between two products or two versions of the same product.

There are several tests for evaluating the user satisfaction. Examples of these are SUMI (Software Usability Measurement Inventory) and QUIS (Questionnaire for User Interface Satisfaction). QUIS (Chin et al., 1988) has 27 items to be rated on a 10 point semantic differential scale. SUMI is a generic usability tool that comprises a validated 50 items questionnaire (KiraKowski, 1996; Macleod et al., 1997). Each item is scored on a three point Likert scale (i.e., agree, undecided, disagree). More recently and due to the rapidly changing patterns and technology of computing today, two new questionnaires are being developed, MUMMS (Measuring the Usability of Multi-Media) to assess multimedia software and WAMMI (Website Analysis and Measurement Inventory) to assess web sites.

For evaluating user satisfaction, we can use existing tests, if they are adequate to the particular case, or we may create a new instrument to fit it. It is important to take into account that the data to be collected should be based on the specified problem statement and test objectives, presented in the test plan.

Another relevant aspect concerning usability testing is when tests should be carried out.

When to carry out usability tests

Usability tests can be carried out at different points in the design and development process (Nielsen, 1993; Preece, 1993; Rubin, 1994; Smith & Mayes, 1996). However, usability testing is most powerful and effective when implemented as part of a product development process (Rubin, 1994).

Even during the conceptualization phase of the educational software, attention should be paid to usability issues. It is important to analyze similar products available in the market place and to interview: experts about the hypermedia structure and teachers that teach the content of the product, and to ask the target users about what they would like to have on the product.

Most of the tests are conducted, during the implementation phase.

Test Plan

The test plan is the basis for the entire testing. It addresses the how, when, where, who, why, and what of the usability test. The test plan describes exactly how you will go about testing your educational software. Under some time pressure of project deadline, there could be a tendency to avoid writing a detailed test plan. But this is a mistake, as Rubin (1994) pointed out.

The test plan may include the following sections: purpose of the test; problem statement or test objectives; user profile; method; task list; test environment and equipment requirements; monitor role; data to be collected and final report (Rubin, 1994).

The number of participants to be used depends on the test selected. For achieving statistically valid results, small sample sizes lack the statistical power to identify significant differences between groups. According to Spyridakis (1992) for a true experimental design, a minimum of 10 to 12 participants per condition must be used. However, Virzi (1990) states that for the purpose of conducting a less formal usability test, recent research as shown that four to five participants will expose 80 percent of the usability deficiencies of a product, which are most of the major problems.

Another relevant aspect of the test plan is the monitor role. It may specify what a monitor will be doing, and under what circumstances the test monitor is intervening.

Conducting a Test

When conducting a test, there are four major components that the monitor cannot overlook: the participant greeting and background questionnaire, orientation, running the test, and participant debriefing (Nielsen, 1993; Rubin, 1994). The monitor or experimenter has the responsibility to make the participants feel as comfortable as possible during and after the test. Moreover, the experimenter should have everything ready before the participants show up.

Ethical aspects of tests

Tests should be conducted with deep respect for the users' emotions and well being. Test participation can be a quite distressful experience for the users, because they may feel a tremendous pressure to perform. Even when they are informed that the purpose of the study is to test the system or the educational software and not the user.

Users' name should be kept confidential. Users can be referred to by an identification number and not by names or initials.

On the following section we describe the usability tests carried out upon the hyperdocument "Cousin Basilio" in order to improve it.

Usability tests conducted to "Cousin Basilio"

The hyperdocument "Cousin Basilio: multiple thematic criss-crossings" was developed in HyperCard in a shell, "Thematic Investigator", created by Rand Spiro and his research group, to facilitate the implementation of Cognitive Flexibility Theory principles (Carvalho, 1999). This shell has been used in the research conducted by Jacobson et al. (1996). As it has been used in United States, we decided to evaluate this hyperdocument with Portuguese students, particularly Humanity undergraduate students, bearing in mind that "usability is measured relative to certain users and certain tasks" (Nielsen, 1993: 27).

We evaluated hyperdocument usability according to the following aspects: (i) icon buttons comprehension; (ii) user' s instruction comprehension; and (iii) user' s degree of satisfaction.

According to Gomoll (1990) and Nielsen (1993) orientations, we invited users to participate with the same type of experience as the user of the hyperdocument will have. For

achieving the aims mentioned above, we created two icon buttons tests (A and B), and we developed an observation grid to indicate user characteristics, user performance, and user degree of satisfaction.

Icon buttons comprehension tests

We conducted two tests (A and B) about icon buttons comprehension according to Nielsen (1993; 1995). Test A is about the association between the icon button and its name, and test B, a little more complex, concerns the association among the icon button, its name and its function. We used a paper-and-pencil evaluation (Rubin, 1994). A card of the hyperdocument was printed, its eight icon buttons numbered, and a list of buttons name was presented (Carvalho, 1999: 409). Users were asked to indicate which name corresponds to each icon. The sample had twenty subjects, 3rd year undergraduate students from the Portuguese-English teaching program. The majority of subjects (75%) correctly established the relation between all eight icons and their names. All subjects associated correctly the icons related to the navigation options (Back, Next, and Return). Two subjects considered that the icon button "T" stands for "Thematic Commentary" instead of "Themes description". This mistake can be easily solved when they interact with the hyperdocument, because "Themes description" appeared in a stack whose name is "Themes". Finally, three other subjects made their associations arbitrarily, because there is no connection at all with the button name and the indicated icon button. Anyway, we may conclude that subjects identified quite well the relation between the icon button and its name.

Test B was carried out by twenty students, belonging to Portuguese and Portuguese-German teaching program. A card of the hyperdocument was also printed but this time they have to combine icon buttons number with its name and a description of its function (Carvalho, 1999: 413). The majority of students (70%) succeeded in the combination of the three aspects of the buttons. All subjects associated correctly four buttons: Notes, Back, Next, and Return.

These tests suggested that navigation icons are intuitive and that attention should be paid to the icons used for "Themes description" and "Thematic commentaries".

One-to-one test

One-to-one test was proposed by Tessmer (1993). It implies an observer and an user. The observer gives some tasks to be performed by the user, and observes its performance. According to the aims defined, quantitative data can be collected, e.g. time necessary to understand the functionality of the document, or qualitative data such as users attitudes to the document (Smith & Mayes, 1996).

The test conducted had the following aims: (i) identifying users difficulties in interacting with the hyperdocument; (ii) verifying if users understand the instructions written in each path; (iii) analyzing windows location and users difficulties; and (iv) evaluating texts size and font according to users difficulties in reading.

This test was structured according to three phases. First, the observer or evaluator characterized the user according to its computer literacy, namely frequency of computer usage and

software used. The second phase focused on user's interaction with the hyperdocument for addressing the four aims stated above. In the third phase, the user is asked about his/her opinion about the document.

Eight undergraduate Humanity students accepted the invitation to participate in the hyperdocument evaluation. Based on data collected, we verified that experienced computer users had more facility in interacting with it. Users with less computer experience considered the document complex, but interesting. Users that had familiarity with different kind of software also felt comfortable rapidly and enjoy the hyperdocument. These results agree with the opinion of Shneiderman (1992) about the importance of user familiarity with the computer system and his/her facility in interacting with the new document.

During the second phase, no difficulties with the icon buttons of "Thematic commentary" and "Theme description" were detected. The subjects considered that the font used on the text "Themes" was small. Moreover, users commented on the difficulty to read the text available on one particular window (thematic commentary) because it partially overlap the novel text, and both had the same line sequence, the same font and size. In what concerns the third phase, i.e., user's satisfaction, they mentioned that they liked it and some of them asked "Why don' t we have documents like this to study?"

According to users' comments and performance several changes were introduced in the hyperdocument. No problems with instructions were detected. We noticed that user interaction depends of their previous computer literacy at beginning, but all of them learn to use it quite easily. Finally, their opinion about the hyperdocument was positive.

Final Remarks

This paper focuses on important aspects of usability testing, which are crucial to the success of educational software. It is impossible to do all usability tests, but we would like to stress that expert evaluation should be always conducted as well as some of the tests for evaluating user performance and satisfaction.

We described how the usability tests helped to improve a hyperdocument to be used for educational purposes. Our experience shows us that these tests are time consuming and demand a meticulous planning. However, achieved results compensate greatly!

References

Baecker, R., Grudin, J., Buxton, W. and Greenberg, S. (1995). *Human-Computer Interaction: Toward the Year 2000*. San Francisco, Ca: Morgan Kaufmann Publishers.

Booth, P. (1989). *An Introduction to Human-Computer Interaction*. London: Lawrence Earlbaum Associates.

Carvalho, A. A. A. (1999). *Os Hipermédia em Contexto Educativo. Aplicação e validação da Teoria da Flexibilidade Cognitiva*. Braga: CEEP, Universidade do Minho.

Carvalho, Ana Amélia Amorim (2001). Princípios para a Elaboração de Documentos Hipermédia. In Paulo Dias e Cândido Varela de Freitas (org), *Actas da II Conferência Internacional de*

Tecnologias de Informação e Comunicação na Educação: Desafios'2001/Challenges' 2001. Braga: Centro de Competência Nónio Século XXI da Universidade do Minho, 499-520.

Chin, J. P., Diehl, V. A. & Norman, K. L. (1988). Development of an instrument Measuring User Satisfaction of the Human-Computer Interface. *Proceedings CHI' 88*, 213-218.

European Usability Support Centres (http://lborl.ac.uk/research/husat/eusc/r_usability_standards.html)

Gomoll, K. (1990). Some Techniques for Observing Users. In B. Laurel (ed.), *The Art of Human Computer Interface Design*. Massachuttes: Addison-Wesley, 85-90.

Grudin, J. (1992). Utility and usability: research issues and development contexts. *Interacting with Computers*, 4, 2, 209-217.

Hix, D. & Hartson, H.R. (1993). *Developing User Interfaces: Ensuring Usability Through Product and Process*. New York: John Wiley & Sons.

Jacobson, M., Maouri, C., Mishra, P. & Kolar, C. (1996). Learning with Hypertext Learning Enviroments: Theory, Design and Research. *Journal of Educational Multimedia and Hypermedia*, 5, 3/4, 239-281.

Kirakowski, J. (1996). The software usability measurement inventory: background and usage. P. Jordan, B. Thomas e B. Weedmeester (eds), *Usability Evaluation in Industry*. London: Taylor & Francis, 169-178.

Macleod, M., Bowden, R., Bevan, N. & Curson, I (1997). The MUSIC performance measurement method. *Behaviour & Information Technology*. Vol.16, 4/5, 279-293.

Marchionini, G. (1990). Evaluating Hypermedia-Based Learning. In D. H. Jonassen e H. Mandl (ed.), *Designing Hypermedia for Learning*. Berlin: Springer-Verlag, 355-373.

Nielsen, J. (1990). Evaluating Hypertext Usability. In D. H. Jonassen e H. Mandl (eds), *Designing Hypermedia for Learning*. Berlin: Springer-Verlag, 147-168.

Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. *Proceedings CHI' 92*, 373-380.

Nielsen, J. (1993). *Usability Engineering*. New Jersey: Academic Press.

Nielsen, J. (1995). *Multimedia and Hypertext: the Internet and beyond*. Boston: AP Professional.

Nielsen, J. & Lyngbaek, U. (1990). Two field studies of hypermedia usability. In R. McAleese e C. Green (eds), *Hypertext: state of the Art*. Oxford: Intellect, 64-72.

Preece, J. (1993). *A Guide to Usability: human factors in computing*. Addison Wesley, the Open University.

Rubin, J. (1994). *Handbook of Usability Testing*. New York: John Wiley and Sons.

Shackel, B. (1990). Human factors and usability. In J. Preece and L. Keller (eds.), *Human-Computer Interaction: Selected Readings*. London: Prentice Hall, 27-41.

Smith, C. & T. Mayes (1996). *Telematics Applications for Education and Training: Usability Guide*. Comission of the European Communities, DGXIII Project.

Spyridakis, J. H. (1992). Conducting Research in Technical Communication: the application of true experimental designs. *Technical Communications*, Fourth Quarter, 607-624.

Virzi, R. A. (1990). Streamling in the Design Process: Running Fewer Subjects. *Proceedings of the Human Factors Society*, 291-294.