Usability Testing and Expert Inspections Complemented by Educational Evaluation: A Case Study of an e-Learning Platform

Andrina Granić and Maja Ćukušić

1Faculty of Science, Teslina 12, Split 21000, Croatia // 2Faculty of Economics, Matice hrvatske 31, Split 21000, Croatia // andrina.granic@pmfst.hr // maja.ckukusic@efst.hr

ABSTRACT

This paper presents a comprehensive usability study conducted within the context of a Europe-wide project. The design of the evaluated e-learning platform is based on an innovative approach to the education of young Europeans by integrating into the curricula of a Europe-wide network of 14 schools different state-of-the-art technologies in e-learning. The evaluation methodology brings together end-user assessments and expert inspections, thus providing a detailed students’, teachers’ and experts’ feedback. User testing integrates six empirical methods into a laboratory-based test. Usability inspection ascertains usability problems by means of recognized heuristics and enables an "educational evaluation" of the platform by means of three sets of criteria. The paper aims to present the effectiveness of the engaged evaluation methods as applied to e-learning platforms. It offers implications from the empirical findings of the user-based methods together with a quantitative and qualitative analysis of the employed inspection methods. Special attention is given to the aspect of educational evaluation. The conducted critical usability examination of a large-scale e-learning system across several countries in Europe revealed which of the chosen assessment methods should be combined to provide constructive and valuable improvement suggestions. A more significant contribution of this research is that the used evaluation approach proved successful, providing some general findings and know-how from the experience and could be reused by other researches because of its thorough structure. As there are relatively few existing accounts of usability assessment in the e-learning context, this paper adds to the body of knowledge.

Keywords

e-Learning, e-Learning system, User testing, Expert inspection, Educational evaluation

Introduction

In the context of the inclusive knowledge society, the role of system interfaces that are more closely tailored to the way people naturally work, live and acquire knowledge is unquestionably recognized as important. In addition, the need for active and accessible learning promotes only the e-learning that engages the users effectively. Nevertheless, despite so much publicity and activity, the progress in the field of e-learning has been relatively slow until recently, when problems were often associated with poor designed e-learning applications cf. (SIGCHI, 2001; Granić, 2008). It seems that too much of the research has been driven by technical possibilities, while paying inadequate attention to the area of application. This issue has been ignored for some time, in the hope that new technologies will somehow resolve the lack of real progress. However, to efficiently communicate the contents and improve the learning experience, interaction mechanisms merit particular consideration. Usability studies in the e-learning field are not very frequent despite the important role that usability plays in the success of every e-learning system. If the interface is not transparent and easy to use, the learners/students concentrate on interaction aspects and not on acquiring content. In addition, it has been claimed that usability assessment needs further consideration of the learning perspective. Namely, the approaches to e-learning usability range from those adapted to e-learning to those applying heuristics without special adjustment to the educational context. Accordingly, as an established set of heuristics and a joint evaluation methodology for e-learning systems do not exist yet, there is obviously a need for further research and empirical evaluation.

The paper reports on a case study of an e-learning platform implemented in the network of fourteen European schools. The contribution of this paper is two-fold. First, it critically examines the usability of a large-scale e-learning system across several countries in Europe. The second contribution of the paper is providing some general findings and lessons learned from the experience. Usability testing, which integrated six empirical methods into a laboratory-based test, was complemented with heuristic inspections. Interface compliance with Nielsen’s (1994) traditional principles was enhanced with experts’ judgment of the system’s "educational evaluation" by means of three sets of criteria: Learning with software heuristics (Squires & Preece, 1999), Educational design heuristics (Quinn, 1996) and Pedagogical dimensions (Reeves, 1994). We expect that this contribution with its general findings and know-how from the experience will facilitate the understanding on how to evaluate and improve the usability of...
e-learning systems based on users’ (learners’/students’ and teachers’) and experts’ feedback. Since there are limited studies in the field, this contribution adds to the body of knowledge.

Related Work

Research in the human-computer interaction (HCI) field has provided numerous principles and guidelines that can steer designers in making their decisions. Although applying good design guidelines alone is a good start, it is no substitute for system usability evaluation. In general, usability is context-dependent and is shaped by the interaction between users, tasks and system purpose. A variety of usability evaluation methods have been developed over the past few decades and most are grouped into usability test methods, user-based involving end-users, and inspection methods engaging HCI experts. Research studies involving different kinds of applications, different user groups and evaluation techniques have been conducted and the need for combining the methods is well understood in the usability field; see e.g., Sears & Jacko (2008).

To analyze usability of interaction mechanisms of e-learning systems, more or less standard assessments and studies have been carried out. Some authors have used traditional usability techniques for e-learning system evaluation and have applied Nielsen’s (1994) heuristics for usable design directly (Rentroia-Bonito & Jorge, 2003). Others have proposed to adopt a checklist of well-established principles shared by many lists of guidelines (Parlangeli, Marchigiani & Bagnara, 1999), or have suggested design heuristics without further adjustment to the e-learning context (Dringus, 1995). A number of authors have argued for more synergistic collaborations between usability and e-learning researchers. Squires and Preece (1999) have made an initial attempt toward integration of usability and learning proposing Learning with software heuristics, a list of guidelines adapted to the context. De Villiers (2004) built further on their criteria, adapting and extending heuristics, not for inspection by experts, but as a basis of a questionnaire survey among learners. Drawing on the considerable research devoted to the usability of performance systems and e-learning design, Mehlenbacher et al. (2005) have outlined a set of twenty-one usability heuristics for evaluating e-learning environments and experiences for the designers. Zaharias and Poylymenakou (2009) have proposed a questionnaire-based usability evaluation method that extends the current practice by focusing also on affective considerations (motivation to learn) that might influence e-learning usability. In order to make use of complementing usability techniques, a relatively small number of studies has been made. eLSE (e-Learning Systematic Evaluation) methodology (Lanzilotti et al., 2006) has been derived from SUE (Systematic Usability Evaluation) methodology, originally developed for evaluating hypermedia systems (Matera et al., 2002). eLSE suggests coupling user-testing and specific inspection activities, which use evaluation patterns that precisely describe the activities to be performed during inspection. Bolchini and Garzotto (2008) have proposed MiLE+ which integrates strategies from various traditional evaluation methods, thus offering an analytical guidance to carry out the evaluation. MiLE+ is an evolution of two previous approaches, SUE methodology and MiLE (Milano-Lugano Evaluation) scenario-driven inspection technique (Triacca et al., 2004).

Despite the undertaken research and efforts, the e-learning field still lacks a widespread culture of usability. Approaches that address both the traditional usability and the pedagogical aspects of e-learning systems in the context of use are still a research issue. Additionally, there is a growing need for thorough usability studies whose results would have an impact on "real" e-learning design and development. An integrated systematic evaluation approach to help with the design and development of cost-effective learner-centered solutions is required. Obviously, there is a clear need for further elaboration and empirical validation even more as neither a recognized set of heuristics, nor a consolidated evaluation methodology of e-learning applications are yet available cf. (Ardito et al., 2006; Zaharias & Poylymenakou, 2009).

E-Learning Platform Description

UNITE (www.unite-ist.org) is a European IST-project with the main goal "to contribute to the improvement of Europe-wide education in secondary schools based on innovative principles in technology, pedagogy and learning scenarios, tested by a well-defined validation framework" (UNITE, 2006). The e-learning platform UNITE seamlessly integrates three distinct technologies including their diverse functionalities into an effective e/m-learning environment: an e-learning portal, an eKnowledge repository and a mobile learning component.
Technically, the system is based on Service-Oriented Architecture (SOA) concepts with two access points: the portal and learning management system are accessed via a web server (http://pilot.unite-ist.org), while the mobile devices connect via a dial-in server (Hornung et al., 2008). From a system design point of view, modularity and re-usability of the system are very important and provide benefit. For easier understanding of our research results, some platform features are elaborated further based on (Kouloumbis, Lu and Wunner, 2007; Lu et al., 2006). The users’ area is a personal space (similar to virtual desktop) incorporating the users’ personal page, personal notes, their journal and agenda, their allocated tasks and personal resource area (Infopool repository). It lists the workspaces the user is a member of and gives access to the messaging environment. The Infopool is a "container" for all e-learning content that can be accessed from workspace areas. The Infopool viewer module provides the common content management facilities while other modules support SCORM-compliant reuse and the editing of the learning courses and its metadata. The reusability of content in Infopool is enabled through metadata searching and tagging (using Metadata editor) while Course editor and Course viewer enable creating, editing and publishing new courses to authorized users. The Mediaboard is a virtual "place" where students and tutors can set up an image/map as the front page of their working space, send text, pictures and audio messages to different locations (i.e., zones) on the image or map. It shares files with the Infopool repository. SMS quizzes and SMS quiz engine support activities that require students to send text messages. These activities can be of two types: students interacting with their tutor or with a computer that sends automated replies. The SMS engine tracks the answers and sends an instant feedback message. PPC Author tool and MyLearning player enable creation and playing of various games and quizzes that can be run on Pocket PCs. Students access the online quizzes using their Pocket PCs though the quizzes can be played offline also. MyLearning player supports different formats of learning materials and tracks a student’s journey through them.

Figure 1 shows snapshots of the interface from a workspace (created for the course "Wonderful world of inventions").
A high-fidelity prototype of the system was released at the end of 2007. Two evaluations were performed on the earlier versions: task-centered walkthrough usability test on the platform mock-up (i.e., low-fidelity prototype) and controlled functionality test. A number of identified problems and bugs with different levels of severity were fixed. However, a large empirical study of the high-fidelity prototype was conducted to fine-tune the system design. This comprehensive and well-documented account of a thorough exercise in usability evaluation of an e-learning platform is presented in the following section. We also offer reflections on what worked and what did not, along with general findings and suggestions for doing similar studies in order to be useful to other researchers in the field.

**Evaluation Approach**

The experimental approach adopted to critically examine and assess the UNITE usability is illustrated in Figure 2. We expected to find different problems because of a wide variation in tasks and different assessment methods applied, both empirical and analytic. A number of problems were identified through testing user tasks in a scenario-based testing. Other problems were detected through tasks mentally simulated by experts from both the HCI and the e-learning field using inspection methods. We assumed that the usability testing complemented with inspections that rely upon experts judging the interface compliance with recognized usability principles along with considerations of educational perspective would provide a more accurate evaluation.

---

**Figure 2. Evaluation approach**

To collect quantitative and qualitative data, a number of measuring instruments were used:

- **multi-choice questionnaire**, aimed at obtaining users’ personal information along with the ones related to their possible prior involvement in the UNITE,
- **memory test**, a questionnaire for measuring the number of successfully memorized system functions,
• **attitude questionnaire** called System Usability Scale (SUS) (Brooke, 1996), a simple standard ten-item questionnaire with five-point Likert scale used for users’ subjective valuation; it is argued that SUS yields the most reliable results across sample sizes (Tullis & Stetson, 2004),

• **semi-structured interview**, an instrument for acquiring further subjective feedback,

• **evaluator’s booklet**, a booklet in which experimenter conducting the assessment procedure took notes, described problems identified and filled in information about accuracy of task completion and time spent on task performance, and

• **evaluation form**, a set of heuristics augmented with auxiliary guidelines related to e-learning systems used in heuristic inspection.

All the aforementioned instruments were used in end-user testing except the evaluation form that was used in expert reviews.

User testing along with thinking aloud session were conducted individually with Internet access and screen capturing software for tracing and recording users' actions and navigation. It was based on criteria expressed in terms of four quantitative measures. The relevant usability goals were established as well (Figure 3).

**Usability Testing**

Individual test sessions were conducted separately from February to March 2008 on the high-fidelity prototype. Note that the prototype was cleared of the bugs identified by the mock-up evaluation and functionality test. The system was implemented in Europe-wide network of 14 schools with 512 students and 46 teachers participating in the implementation. Due to technical limitations, usability testing was conducted in lab environment of 9 schools, involving 47 students (9.18%) and 23 teachers (50.00%). No restrictions were imposed on the involved UNITE user population, so this can be interpreted as one of the study limitations. It is possible that users volunteered or were selected as ones which like or dislike the platform, but there is no way to be sure. Regarding the technical setup, it was difficult to standardize the apparatus on nine different locations, so there could have been some minor deviations of e.g., screen sizes and connection’s speed. Recommendations concerning the computer configuration and browser requirements were given to all the persons in charge of the test. However, it is possible that general system performance might have had a slight impact on the users’ comfort or ability to perform a task, particularly since some of the tasks were dependent on content uploading/downloading.

Regarding the general characteristics of the student sample, 47 individual one-hour sessions were held with an equal number of students from eight different countries (Figure 4a). Almost 60% of students (59.57%) were sixteen and seventeen years old (Figure 4b); 57% of them were male and 43% female. The pre-experiment multi-choice questionnaire provided some information about students’ computer experience and possible prior familiarity with the system. The rollout of the system pilot was completed in late 2007 with the experiments conducted in March 2008. In the meantime, both students and teachers were "exposed" to the system, i.e., they were using it for the learning, teaching and testing purposes. It is rather hard to quantify their exposure since it largely depended on the specific...
scenario they were involved in, but it can be assumed that the exposure time rarely exceeded one hour per week. Users themselves were asked to describe their system foreknowledge. Ten students (21.28%) were characterized as novice users, i.e., they were the ones that had only seen the presentation and had not had a chance to work with the system yet.

Concerning the general characteristics of the teacher sample, 23 individual one-hour sessions were held with an equal number of teachers from seven different countries (Figure 5); 39.13% of teachers were male and 60.87% were female. The average teacher age was 40. Based on teachers’ computer experience and prior involvement in the project, only one teacher was characterized as novice and others as experienced users.

Given the remote collaboration, a usability test booklet containing all the testing materials (i.e., pre/post questionnaires, task scenarios, memory test, attitude questionnaire and evaluator’s booklet) was prepared to standardize the procedures at nine test sites. Work scenarios as sequences of typical tasks and user actions showing the system’s basic functionality were elaborated in order to understand the effect of system design in a sample work situation. Five representative tasks for students and six for teachers were chosen (short descriptions of tasks are given on Figure 6). To check the assigned tasks and time interval, clarity and unambiguousness of measuring instruments and adequacy of hardware and software support, pilot testing was performed. A week before the testing a brief presentation of the evaluation objective along with one-hour lecture on the platform main functionalities was given to all participants.

Each individual testing session started with a short introduction when the experimenter briefed the test user about the assessment experiment and the steps that had to be undertaken. First, the participant was asked to fill in the multiple-choice questionnaire. Note: the user could fill in the required forms directly in the computer (digital form) or could
fill in the paper version. Second, each participant was required to use the platform to perform task-based user testing. The test-user was asked to think out loud while carrying out a set of tasks, particularly to articulate whenever there was a problem. An experimenter took notes and documented all the relevant information in the booklet (the use of a recorder was optional). The booklet included several forms, one for every predefined task. The experimenter enumerated and described the problems detected and further explained them in the problem-description column, additionally identified issues related to where, how and how many times the detected problem occurred. Severity rating on a four-point scale (cosmetic, minor, major or catastrophic problem) was also assigned. S/he then filled in the accuracy of task completion in the respective booklet field (characterized as correct, partial or failure) along with the task completion time that was extracted from the Log Management module and filled in the relevant booklet field. At the end of the experimental session, all the forms were collected. Third, upon completion of tasks from the work scenario, each user had to complete a memory test followed by a satisfaction questionnaire. Fourth, every user participated in a short semi-structured interview in which (s)he expressed her/his satisfaction or dissatisfaction with the system design and offered suggestions for improvement. Additionally, the participants rated and commented on the overall implementation of the UNITE’s concepts. The findings of the individual sites were first compiled locally and then combined, interpreted and presented in an evaluation report that was sent to the development team (Kellner et al., 2008).

Usability Inspection

Heuristic inspection was concurrently performed on the same version of the UNITE system by four evaluators from three countries (the United Kingdom, Latvia, Lithuania) who inspected the platform independently. The recruited evaluators fall into the category of "double experts", the specialists who have experience in both the HCI field and the e-learning systems. The inspection ascertained usability problems by means of Usability heuristic (Nielsen, 1994), judging interface compliance with recognized usability heuristics, and enabled "educational evaluation" of the system by means of three sets of criteria: Learning with software heuristics (Squires & Preece, 1999), Educational design heuristics (Quinn, 1996) and Pedagogical dimensions (Reeves, 1994). The heuristics along with the specific principles for "educational evaluation" focused the evaluators' attention as they worked their way through the system, using their expertise to role-play the behavior of a typical end-user. The evaluators were given a short lecture on the system to be evaluated. Their familiarity with the platform was described as low. The introduction session lasted one hour. A general recommendation was to go through the interface at least twice. The first pass (self-guided exploration) helped the evaluator to get a feel for the "flow" of the interaction and the general scope of the system. To perceive how the system was intended to be used by the end-users, a few representative tasks (teachers’ and learners’) were recommended. The second pass allowed them to focus on specific interface elements as they knew how they would fit into the larger whole; evaluators also used self-exploration.

Overall, experts systematically went through the interface and noted problems that violated the general heuristics and the three sets of criteria for "educational evaluation". Detected problems could equally violate several heuristics. The sets of criteria were randomly chosen. The evaluators tried to be as specific as possible listing each usability problem along with assigned severity rating. From each evaluator, problems were recorded as written electronic reports. For an individual expert, the inspection session lasted about four hours. The debriefing session, conducted in a brainstorming mode, focused on a discussion related to severity ratings of problems and possible redesigns to address the major difficulties of the UNITE interface.

Results and Interpretation of Findings

Analysis of students' feedback

Regarding the accuracy of task completion (user’s effectiveness), as assessed by the experimenter and expressed by a grade in her/his booklet, learners were able to perform 78.3% tasks correctly, 6.8% partially and failed to perform 14.9% tasks. Task completion time (user’s efficiency) was extracted from the Log management module of the system after the evaluation session. The mean time to complete all five tasks was 14.76 minutes. See Figure 6 for mean task completion times per task and Figure 7 for the total time required.
Valuable information was extracted from qualitative material collected during thinking aloud sessions. Based on this feedback, the design team was able to make specific interface adjustments. Here are some examples of what evaluators wrote down:

- User initially didn't enter the topic field - generated error message (only error number!);
- User looked for the reply button at the top;
- User sent a message to non-existing user;
- User initially tried to copy and paste the document within UNITE.

The distribution of scores for memorability (a number of correct answers, from 0 to 10 points for ten questions) ranges from 2.5 to 9.5. On the memory test, novice users scored on average 5.75, while experienced ones 6.58. The scores for users' subjective satisfaction acquired by the SUS can range from 0 (very little satisfaction) to 100 (very high satisfaction). The calculated average of the students' satisfaction score is 59.36. Students' ratings of the platform and the project obtained through semi-structured interviews are shown in Figure 8. To analyze the relations of the listed variables, the Pearson $r$ and Spearman $\rho$ correlation coefficients were used. There is a highly significant correlation between all task completion times: time required to complete a certain task does not depend on the system solely, but it also depends on user's skills (previous experience, reactions, etc).
Highest correlations significant at 1% level are between Task 4 and Task 2 completion time ($r=0.668$) and Task 5 and Task 4 completion time ($r=0.684$). Adding metadata (Task 4) and sending messages to Mediaboard (Task 5) are very specific platform tasks. More significant and positive correlations between various subjective or objective measures of the system perception/performance are listed in Figure 9.

Based on the presented analysis, we conclude that the subjective user evaluation is closely related to the user’s age (experience), the ratings of the interface design and of the overall project concepts. While the SUS score and rating of both interface design and project correlate positively, the SUS score and age correlate negatively, meaning that older learners were less satisfied with the system. The highest relation is between two subjective variables – the rating of the overall concepts and the rating of the interface design. We claim that the success of the project in general depends on the technical platform. Task 3 significantly influences total task completion time. Several severity ratings and accuracy of task completions are also significantly correlated.

**Analysis of teachers' feedback**

Regarding the **accuracy of task completion**, the teachers were able to perform 70.59% tasks correctly, 18.38% partially and failed to perform 11.03% tasks. Mean **task completion time** for six key tasks was 24 minutes (see Figure 6). Follow some of teacher’s thoughts and comments written by the experimenter: **User didn’t post a topic but opened a new forum; A lot of resources were available and he could not decide which of them was the most appropriate one; User created the course but did not understand the content of the xml file.**

The calculated scores of the satisfaction questionnaire show that the average teacher’s **satisfaction** score is 53.15. See Figure 10 for the numeric ratings of the interface design and the project in general (1 being the lowest and 5 being the highest score). The distribution of scores for **memorability** (a number of correct answers; from 0 to 12 points for
twelve questions) ranges from 2 to 12. On the memory test, novice users scored on average 5.5 while experienced ones 7.4 on average.

To analyze the relations of the listed variables, we used the Pearson $r$ and Spearman $\rho$ correlation coefficient. There is an interesting negative correlation of task completion time for task 4 and age ($r = -.0646$, significance 0.002 (N=20). Apart from some obvious correlations (e.g., intensity of the use of the Internet and developed IT skills, the overall rating of the project concepts and the task severity rankings), there are several other significant and positive correlations between various subjective or objective measures of the system perception/performance listed in Figure 9.

![Figure 10. Platform and project ratings, N=23](image1)

![Figure 11. Heuristic violations and related severity ratings of the identified usability problems](image2)
Based on the presented analysis, we conclude that the highest relation is again between the two subjective variables – the rating of the overall project concepts and the rating of the interface design. Satisfaction score, as another subjective variable, depends largely on the rating of the system design. The results from the memory test depend on one particular task (task 6, creating a course) and the system design ratings.

**Expert Reviews**

The results from expert assessments represent a list of problems with references to those usability principles that were violated by the design along with the classification of the problem severity (Figure 11).

Severity ratings indicate whether the problem is superficial (1), minor (2), major (3) or catastrophic (4). However, the overall number of 25 major or catastrophic problems indicates that a lot of work regarding the interface redesign has to be done. Experts express concerns about Infopool, SMS quizzes and Course editor. Severity ratings show equal frequency of both superficial and minor problems on the one hand and major and catastrophic ones on the other. The feedback regarding the second and the third heuristic lists was rather poor. As an argumentation, a quote of one expert is offered: "It was quite difficult to evaluate the system against these two heuristics because they were focused on learner's experience. I would say that overall the system could be used to build and house resources, courses and learner 'experiences' which meet the criteria expressed in the learning design heuristics. Equally, poor use of the system, or the inclusion of poorly designed content, could lead to failure against those same design criteria."

![Figure 12. Platform pedagogical dimensions](image)

Regarding the fourth heuristic that UNITE was evaluated against, the trend-line showed that on Reeves' pedagogical dimensions experts perceived the platform as being more toward the right side of the scale, i.e., the constructivist and
cognitive foundations. In February 2008, we used this methodology to rate the pedagogical dimensions of 14 learning scenarios (Ćukušić et al., 2008). The purpose was to provide a qualitative and graphical comparison of scenarios and to create their "profiles". Namely, Reeves considers that numerical values may cause the evaluator focusing on the quantitative scale itself rather than on dimensions of qualitative aspects. While the aspects of his multidimensional model (see Figure 12) mostly represent a framework for comparative analysis, the continuum could be used for planning purposes, as part of the deliberations on the future design of a learning scenario or an educational program. The purpose of this (second) evaluation, in which the same methodology was used, was to evaluate the platform as a whole. The plotted trend line based on the inspection results is illustrated in Figure 12. The experts perceive the platform to be more toward the right side of the continuum implying that the platform promotes constructivist approach, cooperative learning, intrinsic motivation, flexibility, the acquisition of practical experience, etc. These are the aims of most of today's e-learning systems. The fact that the platform tends to be "on the right" only shows that the system design and implementation were in line with teachers’ requirements. The experts’ opinions related to five selected pedagogical dimensions are offered in Figure 13. The comments are related to the system’s pedagogical philosophy, experiential value, program flexibility, level of learner control, and user activity. As can be seen from the remarks, double-experts commented both on UNITE’s usability and its educational aspects.

![Figure 13. Double-experts’ comments on five selected pedagogical dimensions](image)

**Discussion**

To get a better insight into the obtained correlations, it is useful to visualize some interesting findings. Figure 14 illustrates an individual’s (student’s) IT skills and previous experience with the system in relation to his/her scores on the memory test, subjective satisfaction and total time necessary to complete allocated tasks. Learners’ individual characteristics are shown in the first two coordinates: computer skills (IT skills label) and previous experience with...
the platform (UNITE label) marked from 1 to 4 in the student’s self-assessment (1 denoting "no experience" and 4 "worked with the system for some time"). These two variables could be related to the student results on the memory test (M-total label), his/her subjective satisfaction (SUS label) and the time to complete representative tasks (Task-total label).

It is evident that the users that have lower IT skills and no previous experience with the system have very low scores on the memory test, are not quite satisfied with the platform and need more time to complete the tasks. User’s better performance is correlated with her/his better skills and/or higher experience, thus, as in most complex systems, a learning effect cannot be ignored. However, the system interface should be easy to use and learn, especially for novice users. When considering students, an usable e-learning system is not just a resource with a nice "look & feel", but an application that communicates content and structures the interaction in a way that facilitates the learning experience. Consequently, there remains a challenge to design a system that is transparent and usable for users with different skills, experience and ability. We have presented concrete quantified relations between the above-mentioned variables, pointing out extremely high correlations between all achieved times of task completions. This is reasonable, because the time required to complete a particular task depends not only on the system but also on the user’s skills (previous experience, speed of reaction and the like). The highest correlation coefficient ($r = 0.684$) significant at the level of 1% is the time between the completion of task 5 (sending messages to Mediaboard) and task 4 (adding metadata), very specific for this platform, thus resulting in a positive correlation.

Another interesting positive correlation is between two subjective variables: the evaluation of platform design and the overall project success ($\rho = 0.598$). Users who required a lot of time to complete key tasks (e.g., teachers’ average time was 24 minutes, ranging from 13 to 46) were extremely displeased and frustrated with the system, thus giving to the platform and the project low marks. It is noticeable that the overall success of the project largely depends on the success of the technological platform.
Concerning the presented case study of the UNITE e-learning platform we conclude that although many interface problems were identified by expert reviews, it was the user testing that enabled us to determine which problems actually impeded the users' (students' and teachers') ability in successful task completion. The development team was provided with the detailed information regarding goal achievement and prioritized problems. The decision was made regarding whether or not there was sufficient evidence that the platform had met its objectives. Set quantitative usability goals were an objective that served as an acceptance criteria. The list of the outcomes is given in Figure 15.

![Figure 15. Usability goal outcomes](image)

This is the first critical usability study that has employed user-based methods together with a quantitative and qualitative analysis of a number of "educational" inspection methods. Accordingly, it has made a substantial contribution to the research in the area of usability and educational evaluations of e-learning systems at least in three ways:

- presenting a successful systematic evaluation approach that revealed which of the employed assessment methods should be combined to provide constructive and valuable improvement suggestions;
- providing comprehensive assessment results of a large-scale e-learning system implemented in the network of fourteen European schools, revealing some of the technical and pedagogical issues that could obstruct effective use of the system;
- offering thorough usability examination in educational settings whose results have an impact on "real" e-learning design and development.

The study supported the assertion that we should not rely on isolated evaluations and that expert reviews are not yet a substitute for end-user testing. Actually, those are complementary approaches. Users are oriented toward tasks accomplishment and subjective look and feel of the system design, and hence the results achieved through user testing are appropriate for identification of general usability problems. On the other hand, experts go deeply into the structure trying to identify problems that influence system functions. Therefore, inspection provides a more precise detection of usability setbacks and at the same time offers suggestions for possible solutions. Concerning the employed inspection methods, Nielsen’s (1994) traditional heuristics along with Reeves’s (1994) *Pedagogical dimensions for educational evaluation* provided enough qualitative and quantitative feedback. On the contrary, other two sets of criteria, Squires and Preece’s (1999) *Learning with software heuristics* and Quinn’s (1996) *Educational design heuristics*, showed poor applicability and provided very modest experts’ feedback.

Consequently, there are several important implications on how e-learning systems should be designed and evaluated based on issues identified through users’ scenario-based testing and detected through inspection methods by experts from both the HCI and the e-learning field:

- Expert evaluations were fruitful part of the study that resulted in an exhaustive list of problems relating them to those principles that were violated by the system design, offering also a classification of the problem severity. Contrary to expectations, the feedback regarding *Learning with software heuristics* and *Educational design heuristics* was rather poor. It could be argued that it was quite difficult to evaluate the system against these two heuristics because they focused more on learner's experience.
- Reeves's multidimensional model, the fourth heuristic that the system was evaluated against, proved useful since it provided extensive comments from double-experts on both system’s usability and its educational aspects. Additionally, it was also valuable in our research when several learning scenarios were qualitatively and graphically compared. The same model could be used for planning purposes as part of the deliberations on the future design of a learning scenario or an educational program.
Experts also acknowledged the fact that even if the system is designed to meet the criteria expressed in the learning design heuristics, if it is used improperly, or stores poorly designed learning content, it could lead to failure against those same design criteria.

The study revealed the highest correlation between two subjective variables: the rating of the overall project concepts and the rating of the interface design. Prior research has shown a positive impact of a well-designed technical platform on the success of the e-learning project in general. Our study extends this research by finding that subjective user evaluation is closely related to user's age (experience).

Our experience strongly recommends that, in order to develop user-centered e-learning solutions, is crucial to start validation activities at the early design stage and continue to employ diverse assessment methods throughout the whole development process.

Overall, our findings suggest that there is a value in exploring aspects and strategies for enhancement of "traditional" usability assessment of an e-learning system with educational evaluation. We conclude that accurate assessment is provided by end-user testing complemented with expert inspection that relies upon specialists judging the interface compliance with Nielsen’s recognized usability principles along with considerations of Reeves’ pedagogical perspective. Thus, detailed students’, teachers’ and experts’ feedback from both technical and pedagogical viewpoint could be provided. We conjecture that the effectiveness of the usability testing depends on the experimental design, the chosen tasks and the experimenters, while in the inspection we could speculate on the expertise of engaged double-experts. That's why the results of end-user testing and expert inspections would be far from similar, even if other researchers/teams chose to apply the same integrated systematic approach to evaluate the same e-learning platform, cf. (Molich et al., 2004).

Although traditional heuristics have already been modified/extended and specified to cope with some distinct features of the e-learning applications, the e-learning field and instructional design still lack a widespread culture of usability. The approaches that address both the traditional usability and the pedagogical aspects of e-learning systems in the context of use are still a subject of research. An integrated systematic evaluation approach to help design and develop cost-effective learner-centered solutions is required. We expect that this empirical investigation by itself could provide sufficient motivation for educational system developers and pedagogues to continue to emphasize the integration of usability principles into present e-learning systems. The study offers insights and better understanding on how to evaluate and improve the usability of e-learning systems based on users’ (learners'/student’s and teachers’) and specialists’ feedback. The recommendations deduced from our findings could be useful to other researchers in the field, at the same time aiding in the enhancement of the adoption of usability techniques by designers and project teams as well.

Conclusion

The progress in the field of e-learning has been rather slow, with problems mainly related to poor design of e-learning systems. So far, the development focus has been more on technology aspects rather than on user-centered design issues. Due to the underestimated importance of usability, there are limited studies in the field. This paper reports on a comprehensive usability study that took place within the context of a Europe-wide project. The design of the e-learning platform evaluated is based on an innovative approach – to provide novel services in education for young Europeans. The contribution of the paper is two-fold: a case study of evaluating a large scale e-learning system across several countries in Europe along with providing general findings and lessons. The evaluation procedure includes inspection reviews, judging system interface compliance with recognized usability heuristics and as well as enabling "educational evaluation", and end-user assessments that embody an integration of six empirical methods into laboratory-based usability testing. The research and experimental work undertaken within the context of this Europe-wide project are in line with the growing need to intensify the development of new usability evaluation approaches for e-learning and/or to advance the existing ones.

Acknowledgments

This work has been carried out within the project UNITE 026964: Unified e-Learning environment for the school, partially supported by the European Community under the IST priority of the 6th Framework Programme for R&D. The research has also been supported within the project 177-0361994-1998 Usability and Adaptivity of Interfaces for
Intelligent Authoring Shells funded by the Ministry of Science, Education and Sports (MZOŠ) of the Republic of Croatia.

References


