Systematic Evaluation of e-Learning Systems: an Experimental Validation

C. Ardito°, M. F. Costabile°, A. De Angeli*, R. Lanzilotti°
°Dipartimento di Informatica, Università di Bari, Italy
{ardito, costabile, lanzilotti}@di.uniba.it
*School of Informatics - University of Manchester Po BOX 88 M601QD - UK
Antonella.de-angeli@manchester.ac.uk

ABSTRACT
The evaluation of e-learning applications deserves special attention and evaluators need effective methodologies and appropriate guidelines to perform their task. We have proposed a methodology, called eLSE (e-Learning Systematic Evaluation), which combines a specific inspection technique with user-testing. This inspection aims at allowing inspectors that may not have a wide experience in evaluating e-learning systems to perform accurate evaluations. It is based on the use of evaluation patterns, called Abstract Tasks (ATs), which precisely describe the activities to be performed during inspection. For this reason, it is called AT inspection. In this paper, we present an empirical validation of the AT inspection technique: three groups of novice inspectors evaluated a commercial e-learning system applying the AT inspection, the heuristic inspection, or user-testing. Results have shown an advantage of the AT inspection over the other two usability evaluation methods, demonstrating that Abstract Tasks are effective and efficient tools to drive evaluators and improve their performance. Important methodological considerations on the reliability of usability evaluation techniques are discussed.

Author Keywords
E-learning system evaluation, usability evaluation techniques, controlled experiment.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
E-learning is the most recent way to carry out distance education by distributing learning material and processes over the Internet. Its “any time, any place” nature could be part of a winning strategy for particular needs, such as decongestion of overcrowded education facilities, support for students or teachers who live far away from schools and universities, and adult education.

Making remote educational data and tools available to learners requires considering their different characteristics, such as cultural background, technical experience, technological equipment, and physical/cognitive abilities. Thus, a major challenge for designers and Human-Computer Interaction (HCI) researchers is to develop software tools to engage novice learners and support their learning even at a distance. This could require revising traditional interaction paradigms to provide new flexibility and adaptiveness, to account for the peculiarities of the specific application field. Towards this end, there should be a synergy between the learning process and the student’s interaction with the software. As for any interactive system, the quality of the user interface is a primary requirement. If the e-learning system interface is not usable, the student could spend more time learning how to use the system rather than learning the educational content. Thus, in this particular context, issues of usability will assume greater significance. Besides being usable, an e-learning system must be effective in meeting the instructor’s pedagogical objectives and the student’s needs. Interface elements and functionalities, information quality and pedagogical content should support people to learn in various contexts according to selected pedagogical objectives. As a consequence usability evaluation need to integrate the assessment of pedagogical quality of e-learning.

For the above reasons, the evaluation of e-learning systems deserves special attention, and evaluators need appropriate guidelines as well as effective evaluation methodologies [15]. Unfortunately, the number of studies addressing usability of e-learning systems is relatively small and inadequate to the importance of the issue [11, 13]. Moreover, it is often the case that the evaluation criteria are only vaguely stated [10, 12, 14], so that an actual measurement of the system quality is left to subjective interpretation.
Usability of e-learning systems has been discussed in [1, 7], also introducing eLSE (e-Learning Systematic Evaluation), a methodology aiming at increasing the reliability and the effectiveness of e-learning evaluation by proposing a structured and systematic approach to it. To reach this objective, eLSE proposes the use of an inspection technique, which exploits evaluation patterns, called Abstract Tasks (ATs), in order to drive the evaluators in their inspection activities. For this reason, it is called AT inspection.

This paper presents an empirical validation of the AT inspection for evaluating e-learning systems by comparing it with two other well-known evaluation techniques, namely user-testing and heuristic evaluations. The comparison was based on three major dimensions: effectiveness, efficiency, and satisfaction. Results showed a promising advantage of the AT inspection over the other techniques, demonstrating that Abstract Tasks are efficient tools to drive evaluators and improve their performance.

The paper has the following organization. Next section briefly illustrates the basic usability evaluation techniques and provides the rationale for the AT inspection, which is described in the successive. The section “The Experiment” reports the validation experiment and the section “Conclusion” closes the paper.

**USABILITY TECHNIQUES**

Different methods can be used for evaluating the usability of interactive systems. Among them, the most commonly adopted are user-based methods and inspection methods. User-based methods mainly consist of user testing, in which usability properties are assessed by observing how the system is actually used by some representatives of real users [3]. Usability inspection methods involve expert evaluators only, who inspect the application and provide judgments based on their knowledge and experience [9].

User-based evaluation is generally considered the most complete form of evaluation, because it assesses usability through samples of real users. However, this technique has a number of drawbacks, such as the difficulty to select adequate user samples and to reproduce ecological settings of usage in a limited amount of time. Failures in creating real-life situations may lead to “artificial” conclusions rather than to realistic results. Furthermore, the effort and time which are needed to set up reliable user testing are often conspicuous. A frequently used technique within the user-based methods is the thinking aloud in which users are asked to verbalize their actions during the interaction with the system [3]. This technique offers a window over the users’ mental models, allowing evaluators to detect possible misconceptions about the system and the interface elements which cause them.

Usability inspection methods are more subjective than user-based ones, as they heavily depend on the inspector skills. Their main advantage is related to cost-saving: they “save users” and do not require any special equipment, nor lab facilities [6]. In addition, experts can detect a wide range of problems and faults of a complex system in a limited amount of time. Among the inspection methods, the most commonly used is the heuristic evaluation [9]. It involves a small set of experts inspecting the system and evaluating the interface against a list of recognized usability principles, i.e. the heuristics. Heuristic evaluation is considered a “discount usability” method, very efficient and with a low cost-benefit ratio [9]. Yet, it has some drawbacks, such as its dependence on the inspectors’ skills and experience [4]. Furthermore, the heuristics driving the evaluation are often too generic and not adequate to inform the activities of less experienced evaluators.

In order to overcome these major drawbacks, the AT inspection technique is proposed. It exploits a set of evaluation patterns, called Abstract Tasks (ATs) in order to support the inspectors during their evaluations. In fact, these tasks precisely describe which objects of the application to focus on and which actions to perform during a usability inspection. In this way, even less experienced evaluators are able to produce more complete and precise results. The AT inspection derives from the SUE (Systematic Usability Evaluation) inspection proposed in [8]. During the inspection, the evaluator uses the ATs to perform a rigorous and systematic analysis and produces a report in which the discovered problems are described, as suggested in the AT. The list of ATs provides a systematic guidance to the evaluator on how to inspect an application. Most evaluators are very good in analysing certain features of interactive applications; however, they often neglect some other features, strictly dependent on the specific application category. Exploiting a set of ATs ready for use allows evaluators with limited experience in a particular domain to perform a more accurate evaluation.

The inspection based on AT can be used for evaluating different types of application in various domains, e.g. museum hypermedia [8]. It has been adapted to the e-learning context by defining specific ATs [7]. A controlled experiment, described in [2], has shown that, in the evaluation of multimedia software, this inspection is actually more effective and efficient than Nielsen’s heuristic evaluation [9].

In this paper, an empirical evaluation of the AT inspection to evaluate e-learning systems is presented.

**AT INSPECTION**

According to the eLSE methodology, the AT inspection is the basic activity to perform. The eLSE methodology was proposed to make the usability evaluation of e-learning systems more systematic, efficient, and reliable. Three important characteristics of eLSE methodology are explained below.

1) eLSE couples inspection and user testing, , whenever necessary, to make an evaluation more
The evaluation process starts by having evaluators inspecting the application and identifying possible problems and troubles. The user testing is possibly conducted to validate the inspection findings with real users. Since user testing is designed on the basis of the inspection results, it is better focused and the user resources are optimized. As a result, the overall evaluation is less expensive.

2) eLSE suggests to analyze an application along specific dimensions that address the appropriateness of design with respect to the peculiar nature and purposes of the e-learning applications.

3) The inspection used in eLSE is based on the use of ATs that are specifically defined for e-learning systems. These ATs were defined by considering the literature on e-learning, results of users studies [1], and the experience of usability experts.

ATs are formulated precisely by means of a pattern template, which provides a consistent format and includes five items:

- **AT Classification Code and Title** univocally identify the AT and its purpose
- **Focus of Action** lists the applications objects to be evaluated
- **Intent** clarifies the specific goal of the AT
- **Activity Description** describes in detail the activities to be performed during the AT application
- **Output** describes the output of the fragment of the inspection the AT refers to.

Optionally, a comment is provided which indicates related ATs and usability attributes. Table 1 reports an example of AT; the title is “Availability of communication tools”.

<table>
<thead>
<tr>
<th>QU_01: AVAILABILITY OF COMMUNICATION TOOLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus of action:</strong> communication tools</td>
</tr>
<tr>
<td><strong>Intent:</strong> verify the availability of communication tools</td>
</tr>
<tr>
<td><strong>Activity description:</strong></td>
</tr>
<tr>
<td>‒ Identify the offered communication tools</td>
</tr>
<tr>
<td>‒ Try to communicate with other learners or lecturers</td>
</tr>
<tr>
<td><strong>Output:</strong> a description reporting if:</td>
</tr>
<tr>
<td>‒ The communication tools do not permit to cover all the medial channels</td>
</tr>
<tr>
<td>‒ It is not possible to communicate both with learners and lecturers.</td>
</tr>
</tbody>
</table>

Table 1. Example of AT.

HCI researchers define ATs by observing expert inspectors that performed various inspection, thus capturing evaluators’ expertise. In this way, ATs provide information about application domain, tasks and users, since this information is provided in the AT description.

During the AT inspection, evaluators apply ATs, performing the detailed activity stated in them. At the end of the evaluation session, evaluators provide designers and developers with an organized feedback. The evaluation report describes the detected problems using the terminology provided in the AT for referring to system objects or interface elements, and for describing critical incidents. This standardised language increases the report precision and decreases the possibility of misunderstandings.

The eLSE evaluation methodology proposes an important distinction between the e-learning platform (the container) and the educational modules (the content). The e-learning platform is the interface environment which may offer a number of integrated tools and services for teaching, learning, communicating, and managing learning material. The educational modules are the specific pedagogical content provided through the platform. Usability criteria and ATs for the platform have been defined that differ from that defined for e-learning modules, since different features and criteria need to be considered in these two contexts.

**THE EXPERIMENT**

In order to validate the AT inspection technique, a comparison study involving 73 senior students of a HCI class at the University of Bari in Italy has been conducted. The aim of the experiment was to compare the performance of evaluators carrying out the AT inspection with the performance of evaluators carrying out a heuristic inspection, or conducting traditional user testing.

Following the method adopted in [2], the validation metric was defined along the dimensions of effectiveness, efficiency, and user satisfaction, which correspond to the three principal usability factors as defined in [5]. In the validation metric, effectiveness refers to the number of usability problems discovered by an average inspector and their severity. Efficiency refers to the time expended in relation to the effectiveness of the evaluation. Satisfaction refers to subjective parameters reflecting the evaluation of the technique and the application.

For each evaluation dimension, a specific hypothesis was tested contrasting the AT inspection with the heuristic evaluation, as they are both inspection methods. User testing was used in the experiment as an evaluation baseline, as it is normally believed to be the most effective evaluation method.

- **Effectiveness Hypothesis.** As a general hypothesis, we predicted that the AT inspection should increase the effectiveness of the evaluation as compared to the heuristic inspection. Effectiveness is defined in terms of number of detected problems, their severity, and consistency across different evaluators. The expected advantage of the AT inspection should be related to two factors: (a) its systematic nature which helps the evaluator to identify basic application constituents and
(b) the use of ATs which suggest the activities to be conducted over such application constituents. These factors are likely to reduce the subjective aspect of the heuristic inspection, thus improving the reliability of inspection techniques.

- **Efficiency Hypothesis.** We expected that conducting the AT inspection should take longer than conducting a heuristic inspection, as a rigorous application of several ATs is time demanding. Indeed, the inspectors need to evaluate the application performing each step reported in the AT activity description. On the other hand, the heuristic inspection is less structured and therefore should be faster than AT inspection. Nevertheless, we predicted that the AT inspection should be as efficient as the heuristic inspection, as the higher effectiveness of the technique should compensate for the longer time required by its application.

- **Satisfaction Hypothesis.** The results reported in [2] showed that the AT inspection enhanced the inspectors’ control over the inspection process and their confidence on the obtained results. Thus, we predicted that the satisfaction of inspectors using the AT inspection should be major than the satisfaction of the evaluators that used the heuristics.

In the following sections we describe the experimental method adopted to test these hypotheses.

**Method**

In the following sections we describe the experimental method adopted to test the hypotheses of effectiveness, efficiency, and satisfaction of AT inspection and the results obtained.

**Participants**

The study involved 73 students of the University of Bari in Italy. They participated in the experiment as part of their course-work for an advanced HCI course. All participants had a basic knowledge of usability of interactive systems and of evaluation techniques. They also had some previous experience in evaluating software systems by applying Nielsen’s heuristics. In addition, 25 students of a different course were recruited to act as users in the user testing condition. These students did not have any knowledge of usability.

**Design**

The three usability evaluation techniques (AT inspection, heuristic inspection, and user testing) were manipulated between-subjects. Initially, the evaluators were randomly divided in three groups. Each group was assigned to one of the three experimental conditions. The 23 participants of the heuristic inspection group performed the evaluation applying the “learning with software” heuristics proposed by Preece and Squires [12]. In the user testing group, composed of 25 participants, each evaluator observed a student who performed a set of tasks with the e-learning application while thinking-aloud. Finally, the 25 participants of AT inspection group used the inspection technique with ATs, as proposed by eLSE methodology.

**Procedure**

A week before the experiment, all participants were given a 1-hour demonstration of the application to be evaluated. A few summary indications about the application content and the main functions were introduced, without providing too many details. A couple of days before the experiment, a training session of about one hour introduced participants with the conceptual tools to be used during the experiment. Each group participated in their specific training session.

Data were collected in a group setting, but every participants worked individually. The study consisted of two experimental sessions lasting three hours each. During the first session, participants evaluated the e-learning system applying the technique they were assigned to. The heuristic inspection group was provided with a list of the ten e-learning heuristics [12]. Each participant in the user testing group observed a student interacting with the system to perform seven predefined tasks. The AT inspection group received a list of eight ATs to be applied during inspection (see Table 2). The limited number of ATs was due to time constraints in the experimental session. We selected the most basic ATs, which could guide inspectors in the analysis of the main application constituents of the e-learning system.

<table>
<thead>
<tr>
<th>AT Classification Code</th>
<th>AT Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>QU_01</td>
<td>Availability of communication tools</td>
</tr>
<tr>
<td>QU_02</td>
<td>Graphical interface elements</td>
</tr>
<tr>
<td>QU_27</td>
<td>Availability and quality of the course evaluation tools</td>
</tr>
<tr>
<td>QU_35</td>
<td>Quality of the results of the evaluation tools</td>
</tr>
<tr>
<td>AC_06</td>
<td>Organization of a course</td>
</tr>
<tr>
<td>AC_19</td>
<td>Availability of exercises</td>
</tr>
<tr>
<td>AC_24</td>
<td>Organization of the module pages</td>
</tr>
<tr>
<td>AC_28</td>
<td>Validity of the feedback of the evaluation tools</td>
</tr>
</tbody>
</table>

**Table 2. The eight ATs tested in the validation.**

Participants of the three groups had to find usability problems in the application, and to record them on a booklet which differed according to the experimental condition. For the heuristic inspection group, the booklet included ten forms, one for each e-learning heuristic. The form required information about where that heuristic was violated and a short description of the problem. The AT inspection group was provided with a report booklet including eight forms, each one corresponding to an AT. Again, the forms required information about the violations detected through that AT and where they occurred. In the user testing group, the booklet included seven forms, one for every predefined task. The observer enumerated and described the problems that the user encountered performing the task. At the end of the experimental session, all the forms were collected.
A day after, each evaluator was asked to type their discovered problems in an electronic form. This was required in order to avoid readability problems during data analysis. In the electronic booklet, each participant reported the description of the problem, where it occurred, and how it was found. Finally, the evaluator gave a severity rating to the problem in the scale from 1 (I do not agree that this is a usability problem at all) to 5 (usability catastrophe). At the end of the second session, participants were invited to fill in the evaluator-satisfaction questionnaire proposed in [2]. It combined several item formats to measure three main dimensions: user-satisfaction with the evaluated application, evaluator-satisfaction with the inspection technique, and evaluator-satisfaction with the results achieved.

The Application
The evaluated application was a web-based e-learning platform which provided educational courses on different subjects. The platform permitted to access on-line courses, and to perform auto-evaluation tests or final exams. When a student logged into the e-learning platform, they became part of a Virtual Classroom composed by all the students who were registered to the same course. These students could communicate among them and with a tutor through different synchronous and/or asynchronous communication tools (chat, e-mail, forum). The e-learning platform was composed of 16 web pages having different functionalities.

Participants in the experiment could access a pre-set course on information technology. This course was composed of seven modules, including management of documents, text elaboration, spreadsheets, databases, internet and e-mail. Each module contained about 25 lectures, each one of about 20 web-pages, including explanations, examples, and some practical interactive exercises. At the end of the module there was an auto-evaluation test aimed at verifying the knowledge achieved by the student.

Results
Two expert usability evaluators independently examined all the electronic forms in order to identify single and unique usability problems. The initial inter-rater reliability was .80 and all differences were solved by discussion. This analysis led to the identification of 247 problems and 49 non-problems, or statements which reported not understandable content or unverifiable information. Non-problems accounted for 7% of the statements written by the inspectors applying the heuristic inspection, and 3% of the problems written by the inspectors applying the AT inspection or performing user testing.

Effectiveness analysis
On average, participants reported significantly more problems when applying the AT inspection techniques (mean = 21), than when performing user testing (mean = 9.62) or the heuristic inspection (mean = 12.22). The difference is significant as demonstrated by the results of an Anova with evaluation technique as between-subjects factor ($F_{(2,70)} = 25.45$, $p < .001$). Post-hoc tests applying the Bonferroni’s correction indicated that the effect was due to the higher number of problems reported in the AT inspection condition, while there was no significant difference between heuristic inspection and user testing.

Usability problems were classified into five categories according to their cause. Table 3 reports frequencies and percentages of these categories in the three experimental conditions.

- **Content** included negative statements on the quality of the information provided by the system in terms of subject-matter clarity and completeness, as well as poor information architecture.
- **Graphical Design** regarded adverse comments on aesthetic aspects of the interface, including the use of colours, text layout and fonts, as well as poor animations and pictures.
- **Technical Problem** covered issues of pages visualization, compatibility of the system with the browser, and downloading time.
- **Feedback** included negative statements addressing the communication between the user and the interface.
- **Navigation** covered problems related to the task of moving within the system. It referred to the appropriateness of mechanisms for accessing information and for getting oriented in the system.

The inspectors of the three groups highlighted the difficulty of interacting with the communication tools: 60% reported problems using the forum and the 19% with the chat. They reported that it was hard to understand not only how to find them but also how to use them. Other frequent problems addressed poor graphical design complaining about use of colours, font inconsistency, inappropriate use of flashing icons. As regard to system feedback, participants complained about lack of tooltip. Other remarks referred to difficulties in navigation and orientation within the system, for example, the participants had difficulties in finding additional educational material related to a specific lecture.

<table>
<thead>
<tr>
<th>Category</th>
<th>Heuristic Inspection</th>
<th>AT Inspection</th>
<th>Direct Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
</tr>
<tr>
<td>Content</td>
<td>19</td>
<td>17</td>
<td>43</td>
</tr>
<tr>
<td>Graphical Design</td>
<td>25</td>
<td>22</td>
<td>49</td>
</tr>
<tr>
<td>Technical Problem</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Feedback</td>
<td>48</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>Navigation</td>
<td>19</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>100</td>
<td>161</td>
</tr>
</tbody>
</table>

Table 3. Frequency and percentage of usability problems classified by categories and experimental conditions.

To analyse the diagnostic power of the three evaluation techniques in further details, the severity rating attributed
by the expert evaluators to each individual problem was entered as dependent variable in an Anova with technique (3) and problem category (5) as between-subjects factor. Both the main effects and their interaction were highly significant, namely Technique \( F(2,1032) = 6.44, p < .01 \), Category \( F(4,1032) = 48.29, p < .001 \), and Technique * Category \( F(8,1032) = 3.42, p < .01 \). As can be seen in Figure 1, user testing is particularly useful in evincing serious problems related to navigation and graphical design, whereas AT tended to elicit more serious problems related to the educational content. The heuristic inspection consistently scored lower in almost all dimensions.

Six problems were given a 5 in the severity rating (usability catastrophe). They regarded (a) the lack of mechanisms to signal the user position during the navigation leading to disorientation; (b) the inconsistent usage of ambiguous icons and field captions to access the information; (c) a system crash which regularly occurred whenever the user performed a specific command combination; (d) a difficulty in interacting with some system functionalities, such as the lack of text-editing functionalities when performing an exercise or writing a message in the forum. Problems (a) and (b) are of primary importance in e-learning systems and were identified by 9 participants in the heuristics condition; 7 in the user testing condition, and 5 in the AT condition. Problems (c) and (d) were consistently missed by all participants in the AT condition and this may be due to the lack of specific ATs focusing on them. But, they were discovered by 2 participants in heuristic inspection and 5 in user testing.

Efficiency analysis
The distribution of evaluation time was highly skewed. Consequently this variable has been analysed by non-parametric statistics. The Kruskal Wallis test reported a significant difference in evaluation time due to the technique applied (\( \chi^2 = 59.52, p < .001 \)). As expected, participants applying the AT technique required the longest time to carry on their task (mean = 177 minutes; standard error = .46). Participants applying the heuristics worked on average for 137 minutes (standard error = 3.76). Unexpectedly, instead, user testing was the fastest method (mean = 108 minutes; standard error = 1.81). In reading this last datum, one should take into consideration that the user testing average reflects the time needed to test 1 user.

An overall efficiency index was computed, which reflected the number of problems each participant found in 10 minutes. The Anova indicated that overall efficiency is affected by experimental condition, \( F(2,70) = 3.80, p<.05 \). Post Hoc comparisons (Bonferroni’s correction) showed that AT was the most efficient technique (mean = 1.19 problems in 10 minutes, standard error = .08). There were no differences between heuristics and user testing, in the order mean = .91, standard error = .08 and mean = .90, standard error = .08 (Figure 2).

Satisfaction analysis
Satisfaction with the technique applied in the evaluation was assessed by 11 items of a semantic differential scale. The reliability analysis returned an unsatisfactory value (\( \alpha = .73 \)) suggesting that the scale may be composed of separate dimensions. A factor analysis confirmed the existence of 3 dimensions, explaining 52% of the variance. The first dimension reflects the reliability of the technique, the second factor addressed its pleasantness and the last one the easy of use. Three indexes were computed averaging scores to the items with a loading superior to .35 on one and only one factor.

Average values of the three factors are illustrated in Figure 3. These three indexes were entered as dependent variables in a multivariate analysis of variance with technique as the between-subjects factor. The multivariate test indicated a tendency for condition (Willk’s Lambda \( F(6,134) = 2.05, p = .06 \)), suggesting that overall these three evaluation dimension changed according to the technique applied by the participants. The analysis of the univariate effects suggested that this is entirely due to the dimension of pleasantness (\( F(2,72) = 5.39, p < .01 \)). Post-hoc test following the Bonferroni method indicated that user testing is evaluated as the most pleasant technique (p < .05) with no differences between the other two conditions.
Satisfaction with the evaluated application was assessed by a 12-item semantic differential scale, which showed high reliability at the Cronbach test (α = .85). An overall index was computed by averaging scores to individual items, and entered as dependent variable in an Anova with technique (3) as between-subjects factor. Results indicated a significant effect for technique $F_{(2,69)} = 3.93$, $p < .05$. Post-hoc analysis applying the Bonferroni method, demonstrated that the effect is due to the difference between participants using AT inspection and participants applying heuristic inspection ($p < .01$). Surprisingly, after applying AT inspection participants were more positive in evaluating the application (mean = 4.87), than participants in the heuristic inspection condition (mean = 4.32). Participants in the user testing condition expressed an intermediate evaluation (mean = 4.56).

![Figure 3. Average of ease, reliability, and pleasure evaluations as a function of the techniques.](image)

**Discussion**

This paper reports important findings on the quality of different evaluation methods for e-learning systems. The validation study has supported both the effectiveness and the efficiency hypotheses, whereas it did not support the satisfaction hypothesis. Indeed, using the AT inspection, participants were able to find a greater number of usability problems, which were also more serious than those found by inspectors using the heuristic inspection. Finally, we have demonstrated that the AT inspection is also more efficient than the heuristic evaluations, as it allowed participants to find more problems in the same time.

Overall, AT out-performed our expectations in comparison with user testing, which is generally regarded as the best option for usability assessment. The AT inspection was more reliable and helped finding more problems in the same amount of time needed to test one individual user. The study also revealed that different techniques addressed different type of problems. User testing and heuristic inspection helped to highlight problems common to all interactive systems, whereas the AT inspection focused also on specific problems of e-learning. These findings confirm the distinct characteristics of AT that specifically addresses the e-learning domain, provided that appropriate ATs are defined. This is an obvious but important limitation of the AT inspection. The study demonstrated that the evaluators tend to miss serious problems whenever they are not directly addressed by the ATs utilised during the evaluation. The effect of this limitation is reduced by providing a longer time to the inspectors for their evaluation.

The better performance achieved by participants using AT inspection was not mirrored by their subjective evaluations. Indeed, the only difference in evaluator satisfaction among the three techniques revealed an advantage of user testing as regards to pleasure in using the technique. This can be explained by the social nature of user testing.

**CONCLUSIONS**

One present goal of Human-Computer Interaction researchers and developers is to create software tools that support people to learn in an effective manner the material available online. The study reported in this paper has provided some answers about the effectiveness, efficiency and satisfaction of different evaluation techniques, namely heuristic inspection, user testing and AT inspection. The experiment confirms our general hypothesis of an increase of the overall quality of inspection when Abstract Tasks are used. More specifically, the AT inspection was found to increase the evaluation effectiveness and efficiency with respect to both the heuristic inspection and user testing.

Our work objective is to identify evaluation techniques specific for the e-learning domain. The number of studies referring to usability of e-learning systems is still small and specific evaluation criteria are vaguely stated [10-14]. This paper shows that AT inspection is capable to address specific issues of e-learning better than other techniques such as a specialised heuristic evaluation and user testing.

The study also unveiled an often disregarded problem with usability evaluation techniques in general, namely its lack of reliability. The concept of reliability refers to the extent to which a measurement can be replicated. It concerns the problem of generalising results of a study to other users, conditions, times, and places. It means consistency and is a logical precondition to scientific validity. In our validation experiment, we found that different evaluators tended to produce very different usability report. Even using the AT inspection, which resulted to be the most reliable technique, the internal consistency among evaluators’ reports was less than 15% as regards to the more serious problems affecting usability. In the study presented in this paper, the problem of reliability may have been increased by the limited time available to our participants to perform their evaluation. With more time they would probably converge on a larger set of common problems. Another aspect that is worth mentioning is that the participants were novice evaluators. It is remarkable that, despite their limited experience, they were nevertheless able to obtain significant results with the AT inspection. To better assess the value of the AT inspection, we plan to conduct in the future a similar experiment involving professional evaluators.
REFERENCES


