Lanzilotti R., Ardito C., Costabile M. F., De Angeli A. (2010). Do patterns help novice evaluators? A comparative study. *International Journal of Human Computer Studies*, Vol.69(1-2), January, 2011; pp. 52-69, ISSN:1071-5819

Do patterns help novice evaluators? A comparative study

R. Lanzilotti°, C. Ardito°, M. F. Costabile°, A. De Angeli*+

°Dipartimento di Informatica, Università di Bari, Italy
{lanzilotti, ardito, costabile}@di.uniba.it

*Manchester Business School - The University of Manchester, UK
Antonella.de-angeli@manchester.ac.uk

+Department of Information Engineering and Computer Science, University of Trento, Italy
antonella.deangeli@disi.unitn.it

ABSTRACT

Evaluating e-learning systems is a complex activity which requires considerations of several criteria addressing quality in use as well as educational quality. Heuristic evaluation is a widespread method for usability evaluation, yet its output is often prone to subjective variability, primarily due to the generality of many heuristics. This paper presents the Pattern-Based (PB) inspection, which aims at reducing this drawback by exploiting a set of evaluation patterns to systematically drive inspectors in their evaluation activities. The application of PB inspection to the evaluation of e-learning systems is reported in this paper together with a study that compares this method to heuristic evaluation and user testing. The study involved 73 novice evaluators and 25 end users, who evaluated an e-learning application using one of the three techniques. The comparison metric was defined along six major dimensions, covering concepts of classical test theory and pragmatic aspects of usability evaluation. The study showed that evaluation patterns, capitalizing on the reuse of expert evaluators know-how, provide a systematic framework which reduces reliance on individual skills, increases inter-rater reliability and output standardization, permits the discovery of a larger set of different problems and decreases evaluation cost. Results also indicated that evaluation in general is strongly dependent on the methodological apparatus as well as on judgement bias and individual preferences of evaluators, providing support to the conceptualisation of interactive quality as a subjective judgement, recently brought forward by the UX research agenda.

Keywords: usability evaluation techniques, e-learning evaluation, evaluation patterns

1. Introduction

Since the term usability entered common usage in the early 1980's, different approaches and a set of methods have been proposed to evaluate interactive systems (for example, Nielsen and Mack, 1994; Dix et al., 2003; Preece et al., 2007). The objective of this complex methodological apparatus was to propose effective solutions to measure the quality of interactive systems, identify problems, and suggest remedial actions. Despite several studies have been performed to compare different approaches to usability evaluation, in practice the selection of a specific method is often based on considerations of costs and available resources (Hartson et al., 2003; Ssemugabi and de Villiers, 2007; Law et al., 2009). Less emphasis is paid to the power of different methods, in terms of essential measurement properties, such as validity and reliability (Gray and Salzman, 1998; Wixon, 2003; Blandford et al., 2008).

Yet, there is evidence that evaluation results are affected by judgement bias and individual preferences, leading to random or even systematic errors. It was reported, for example, that evaluators applying the same method are likely to highlight different sets of problems or to attribute variable importance to the same problem (Doubleday et al., 1997; Hertzum and Jacobsen, 2003; Liljegren, 2006; Ling and Salvendy, 2009; Tan et al., 2009). Similarly, different techniques can lead to variable results, independent of system features, evaluator expertise and problem severity (Ssemugabi and de Villiers, 2007; Frøkjær and Hornbæk, 2008).

This paper contributes to the debate on the relative power of different methods for evaluating *e-learning systems*. This is a challenging domain where evaluation requires to take into consideration several factors besides traditional usability (Squires and Preece, 1999; Notess, 2001; Meira and Peres, 2004; Lanzilotti, 2006). An obvious factor refers to educational quality, where responses such as engagement, motivation, and feeling of control become of fundamental importance. As emphasised by Squires and Preece (1999) "there is a need to help evaluators consider the way in which usability and learning interact". Some guidelines and evaluation criteria have been proposed but they are still vaguely stated so that an actual measurement of quality is left to subjective interpretation and skills (Parlangeli et al., 1999; Wong et al., 2003; Hornbæk, 2006). The Pattern-Based (PB) inspection is proposed to reduce such drawbacks; it exploits a set of evaluation patterns to systematically drive inspectors in their evaluation activities.

The PB inspection is a general method, applicable to the evaluation of any interactive systems, provided that a proper set of evaluation patterns is defined. In this paper, the application of the PB inspection to the evaluation of e-learning systems, and a study that compares it to heuristic evaluation and user testing are described. Results showed that evaluation patterns provide a systematic framework which drives the evaluators in their analysis, reducing reliance on individual skills, increasing inter-rater reliability and output standardization, addressing a larger set of different problems, and decreasing the evaluation cost. It also highlighted the possible risk of attentional fixity by which evaluators may miss major problems if they are not directly addressed in the set of evaluation patterns available for the inspection.

The paper is organised as it follows. Section 2 covers related works describing different approaches to usability evaluation and reporting on comparative studies investigating their relative strengths and weaknesses. Section 3 presents the pattern-based evaluation approach and how it is applied to evaluate e-learning systems. Section 4 describes the comparative study, proposing also a comprehensive metric for assessing the quality of evaluation techniques. Section 5 and 6 report the results and discuss them. Section 7 concludes and provides suggestions for future research.

2. Related works

A large set of evaluation techniques exist in the literature. Preece et al. (2007) clustered them in three main approaches: usability testing, analytical evaluation and field studies. In the 1980s, usability testing was the dominant approach and still remains important, although, over the years, analytical evaluations and field studies have grown in prominence. Usability testing concerns the analysis of users' performance on the tasks for which the system is designed (Preece et al., 2007). This approach has the potential to provide reliable results since it involves samples of real users. However, reproducing realistic situations of usage in a laboratory is difficult, e.g., selecting a representative sample of users and tasks, training users to master advanced features of the system in a limited time period, or weighting the effect of important contextual factors on their performance (Lim et al., 1996). The cost and time needed to set up usability testing may also be considerable. A frequently used technique in usability testing is *thinking aloud,* which requires users to speak out loud their thoughts while performing tasks. Evaluators detect problems by observing the user behaviour and listening to their thoughts, so that they can follow users' reasoning.

The analytical approach includes inspections and the application of formal models to predict users' performance. A common inspection method is *heuristic evaluation* (Nielsen, 1993). It involves experts who inspect the system and evaluate the interface against a list of usability principles, i.e., the heuristics. The main advantage is related to cost-saving: they "save users" and do not require special equipment or lab facilities (Jeffries and Desurvire, 1992). In addition, experts can detect a wide range of problems of complex systems in a limited amount of time. The main drawback of such a technique is the dependency on the inspectors' skills and experience, as heuristics are often generic and underspecified (Doubleday et al., 1997; Law, 2007). To counteract this problem, *checklists* have been proposed. They simplify heuristics in specific questionnaire items which have to be scored by evaluators.

Field studies differ from the other evaluation approaches because they are conducted in natural settings. Their aim is to understand what users do naturally and how technology impacts on them. They are useful for identifying opportunities for new technology, eliciting requirements, deciding how best to introduce new technology, and evaluating technology in use. Evaluating usability in the field is difficult, due to the complexity of the environment and the activities to be observed, and to the large amount of data to be analysed (Pascoe et al., 2000).

Section 2.1 reports results of studies which compared analytical evaluation with usability testing, while Section 2.2 focuses on evaluation of e-learning systems.

2.1 Comparative studies

Table 1 summarizes the main variables analysed by 10 studies which have compared the analytical approach with usability testing, and their results. The symbol ">", in the results column, indicates an advantage of a technique over another, whereas "=" indicates no difference. It is interesting to notice that, despite emerging critiques on the validity of problem count as a measure of quality (Gray and Salzman, 1998; Blandford et al., 2008; Cockton and Woolrich, 2009), this variable was systematically used by all studies. Less attention was devoted to the assessment of the method scope, a variable which describes 'what *kinds* of insight' a specific method produces (Blandford et al., 2008). This measure can inform a deeper understanding of the degree of complementarity, contradictoriness and overlap between the results achieved by different techniques, but it was not systematically considered in the studies surveyed, and whenever possible it has been inferred by comments and discussion provided by the authors.

The analysis of Table 1 shows contradictory results with regard to the relative power of different evaluation techniques. Four studies proved an advantage of the analytical approach over usability testing in the number of problems found, while evincing an advantage of usability testing over the analytical approach in terms of problems severity. Doubleday et al. (1997), for instance, reported that HCI experts performing heuristic evaluation identified more and different types of usability problems than usability testing supported by thinking aloud. Moreover, users and experts described usability problems in different ways; specifically, "…the end user indicates the *symptom* of the problem …, HCI expert tries to identify the *cause* of the problem". Similar results are reported in (Jeffries et al., 1991; Law and Hvannberg, 2002). Frøkjær and Hornbæk (2008) found a similar effect comparing heuristic evaluation, cognitive walkthrough, and thinking aloud against a new analytical technique, called Metaphors Of human Thinking (MOT), in which the user interface is inspected using metaphors of habits, stream of thoughts and knowledge. MOT identified the same number of problems as heuristic evaluation, but more problems than cognitive walkthrough and thinking aloud. However, evaluators considered problems found applying MOT to be less serious than problems found by thinking aloud.

Table 1. Summary of studies comparing analytical evaluation with usability testing, ordered by publication date

Contrary to this trend of results, three studies suggested that usability testing was more accurate in finding problems than analytical methods, even if these latter methods were more cost-effective. For example, Karat et al. (1992) reported that thinking aloud identified a largest number of usability problems, followed, in the order, by team walkthrough and individual walkthrough. However, the authors stressed that thinking aloud was heavily dependent on the evaluators' ability to conduct the test, their experience with the system and HCI knowledge necessary to interpret the user's behaviour and identify usability problems. They also acknowledged that walkthrough was a good alternative when resources are limited, particularly at the early phases of the development cycle. Similarly, Liljegren (2006) demonstrated that thinking aloud is the most effective method in usability evaluation, followed by cognitive walkthrough, heuristic evaluation and hierarchical task analysis. Nielsen and Phillips (1993) compared the results of heuristic evaluation, GOMS analysis and direct observation on two user interfaces. The study took into consideration several factors in order to evaluate the cost of each method, including the cost needed to build a running prototype of sufficient stability. Three different heuristic evaluations were considered: (*a*) cold condition, i.e., evaluators inspected the specifications of the two interfaces; (*b*) warm condition, i.e., evaluators inspected a running prototype of one interface and the specification of the other interface; (*c*) hot condition, i.e., evaluators inspected the running prototype of both interfaces. The GOMS analysis was performed on specifications, while usability testing was performed on running prototypes. Usability testing was found to be the best method, even if it was more expensive than cold heuristic and somewhat more expensive than GOMS. The cost of heuristic evaluation increased in the hot condition, leading the authors to recommend usability testing when running interfaces are available. The study further indicated that inspections are particularly useful at the early phases of the development cycle.

Finally, three studies reported no differences between the analytical approach and usability testing. For example, comparing heuristic evaluation with user observation supplemented by data log, user diaries, questionnaires and interviews, Steves et al. (2001) found no difference in terms of number of problems and their severity. Heuristic evaluation required less time and effort to highlight many problems that were also found by observing users in real work situations. The authors concluded that each technique has particular strengths, they are complementary and work well in combination. The analytical approach is best suited to find early and major usability problems and user testing to reveal contextual issues. Similar results and conclusions are reported in (Hornbæk and Frøkjær, 2005; Tan et al., 2009).

To conclude, previous studies provided contradictory findings with regard to the relative power of different evaluation techniques in terms of problem count, severity rating and time requirements. They suggest that different techniques applied to the same technological artefact tend to elicit different results, thus challenging the idea of usability as a quantifiable and objective property of an interactive system (De Angeli et al., 2009). Finally, they indicate that different techniques have specific strengths and weaknesses, and therefore should be used in combination.

2.2 Evaluation of-learning systems

E-learning evaluation deserves special attention, since it is not only sufficient to ensure that an e-learning system is usable, but pedagogical qualities should also be taken into consideration (Squires and Preece, 1999; Notess, 2001; Meira and Peres, 2004; Lanzilotti, 2006). Specific evaluation methodologies must therefore be defined. Formerly, some authors proposed that general usability heuristics could also be applied to e-learning systems (Schwier and Misanchunk, 1993). On the contrary, Squires and Preece (1999) argued that it was indispensable to consider socio-constructivist principles and proposed the "learning with software" heuristics. They included specific principles such as match between designer and learner models, navigational fidelity, appropriate levels of learner control, strategies for cognitive error recognition, and match with the curriculum. Similarly, Quinn et al. (1997) proposed a methodology that takes into account both design factors and acceptance factors. Design factors comprise instructional goals, instructional content, learning tasks, learning aids and assessment, whereas acceptance factors include motivational factors, level of active participation entailed, quality of learning support, and user satisfaction.

Several c*hecklists* have been proposed to simplify the evaluator's task (Gerdt et al., 2002). Ravden and Johnson' checklist (1989) emphasized usability qualities, but it did not address pedagogical issues. On the contrary, the Delta checklist paid attention to cognitive and pedagogical issues (Delta, 2002). Similarly, the Learning Technology Dissemination Initiative (LTDI) checklist covered several aspects of learning with a structured questionnaire (LTDI, 2002), including quality of the interaction and information presentation, and pedagogical issues related to matching strategies with objectives or assessment. However, it did not precisely address technological factors and included some underspecified questions. The TUP (Technology-Usability-Pedagogy) model tried to overcome the drawbacks of other checklists by concentrating on technological, usability and pedagogical issues (Gerdt et al., 2002).

Mendes et al. (1998) published a set of metrics to measure some features of e-learning systems, which are more related to the designer's point of view rather than to the user's point of view. Such metrics refer to maintainability, reusability, application structure, etc., and have been defined by using the Goal-Question-Metric approach (Basili et al., 1994), well known in software engineering. Ng et al. (1999) proposed a "hypertext structure measurement system", based on metrics, to help both educational designers and users to analyze e-learning systems. In particular, they identified four main metrics, related to the hierarchical organization of the teaching units, the clustering of different type of documents (e.g. exercises, evaluation tests, external documents, etc.), the guidance provided by exercises/tests, and the help provided.

Formulating more specific heuristics and/or checklists, or defining metrics, is not enough, since they are not operational and their application still requires skilled inspectors to be able to carry out the specific evaluation activities, having knowledge not only of human factors but also of the application domain, the users and the tasks they perform. Hence, evaluators need tools to support them producing more complete and objective outcomes.

3. Pattern-Based inspection

The Pattern-Based inspection (or PB inspection) described in this paper aims to support the work of evaluators during the inspection, by providing structured guidance in the form of *evaluation patterns*. The concept of pattern was originally introduced by Christopher Alexander within the domain of architecture and urban planning, as a cognitive tool to capture human expertise and to make it reusable (Alexander et al., 1977). Patterns have been used also in the design of computer systems (Gamma et al., 1995; Juristo et al., 2007), specifically in hypertext design (Garzotto et al., 1994; Bernstein, 1998), interaction design (Borchers, 2001; Tidwell, 2005), e-learning systems design (Avgeriou et al. 2003; Dimitriadis et al., 2009): they intend to help designers by providing indications on how to manage specific aspects of a design. For instance, within the context of interaction design, Tidwell (1999) describes a design pattern that suggests artefacts that can make navigation easy, convenient, and psychologically safe for the user; examples of such artefacts are the home button in a web application and the undo feature. A design pattern is composed of several items that give the designer indication on the problem it addresses, the solution it proposes, the context in which it can be applied, etc.

Looking at the use of patterns in the literature, it emerges that the two main features of patterns are (Borcher, 2001): 1) a uniform structure and format; and 2) an effective way to organize complex information according to a combination of elements (the pattern items). This information captures a certain expertise and makes it available to other people. The original Alexander's patterns consisted of the same items, presented in the same sequence and form (Alexander et al., 1977). Each item can be more detailed, as the "context" item in the design patterns proposed in (Avgeriou et al., 2003), or more concise, as the patterns described in this paper. While the value of design patterns is to support designers, the value of evaluation patterns is to support evaluators performing the usability inspection of computer system.

The idea of evaluation patterns was originated by the consideration, reported at the end of Section 2, that inspections, based in heuristics, guidelines, checklists, are highly dependent on evaluators skills (Jeffries et al., 1991; Doubleday et al., 1997; Kantner and Rosenbaum, 1997; Ling and Salvendy, 2009). High professional and experienced evaluators are not affordable by small companies developing ICT systems. The only way to make usability evaluation possible is to train people in the company to perform usability inspection. Evaluation patterns are defined to provide support primarily to novice and not professional evaluators; the rationale is that they capture the expertise of skilled evaluators (i.e., their behaviour in conducting an inspection), and express it in a precise and understandable form, so that this expertise can be reproduced, communicated, and exploited (Nanard et al., 1998). The evaluation patterns used by the PB inspection indicate which are the critical aspects of the application to look for, and which actions to perform during the inspection in order to analyse such aspects. Another advantage of the general concept of pattern is that they supply a common language to the community (Tidwell, 2005). The terminology adopted in the evaluation patterns is used by inspectors for reporting problems, thus the resulting evaluation reports are more consistent and easier to compare. Summing up, the proposed evaluation patterns present several advantages: a) they incorporate usability knowledge and best evaluation practices; b) they enforce standardization and uniformity of evaluation reports; c) they provide information about the application domain, tasks and users.

The PB inspection is illustrated in this paper by applying it to the evaluation of systems in the e-learning domain. It is part of the *eLSE* methodology (e-Learning Systematic Evaluation), which combines inspection and usability testing to achieve more reliable results (Lanzilotti, 2006; Lanzilotti et al., 2006; Costabile at al., 2007). In this methodology, the PB inspection assumes a central role: the evaluation process first applies the PB inspection; occasionally, when more information by the users is needed, user-testing is carried out.

Evaluation patterns addressing the overall quality of e-learning systems have been developed by an iterative approach. Since the main purpose was to capture the expertise of professional evaluators, we started by observing such evaluators at work, focusing on their main activities. We also observed teachers and students using e-learning systems, reviewed e-learning literature, and performed several brainstorming sessions with professional evaluators and e-learning experts. A group of usability experts and e-learning experts analysed all the gathered information and structured it into an initial set of evaluation patterns. These patterns were tested through pilot studies asking novice evaluators to use them and provide comments about their clarity, utility, guidance, etc. Based on these comments, the patterns were refined iteratively. They are systematically formulated by means of the following template, which provides a consistent structure:

- *Classification Code and Title,* which identify the pattern, and succinctly communicate its scope;
- *Focus of Action,* which shortly describes the context to which the pattern applies by listing the application components to be evaluated by it;
- *Intent,* which illustrates the problem addressed by the pattern clarifying the specific goals to be achieved through its application;
- *Activity Prompts,* which prompts the activities to be performed by evaluators during the pattern application;
- *Output,* which suggests a format and a standardised terminology for reporting the results of the inspection.

Evaluators choose specific evaluation patterns to be used during the inspection by reading *title, focus of action* and *intent*. Special attention is devoted to select an appropriate title, so that evaluators can quickly understand if that pattern is worth using in their evaluation. Then, they perform the activities suggested by the *activity prompts* and report their finding according to the *output*.

The first three items of the pattern provide information that can be somehow considered similar to guidelines for inspectors. The drawback of guidelines is that they are not operational, they can help experienced evaluators but they do not provide enough support to novice ones, who still have difficulties in performing the inspection. The remaining two items actually overcome this drawback: the *activity prompts* suggest which actions novice evaluators have to carry out in order to perform an accurate inspection; the *output* indicates how a possible problem has to be reported and the terminology to be used, so that the precision of the evaluation report increases, limiting the risk of misunderstandings and providing reports that are easier to compare.

The study reported in this paper demonstrates that, by exploiting evaluation patterns, less experienced evaluators are able to come out with more complete and precise results. This work is grounded on previous research on the evaluation of hypermedia systems (Matera et al., 2002; De Angeli et al., 2003), and confirms that evaluation patterns provide a systematic framework useful in the evaluation of interactive systems.

We have defined 69 evaluation patterns, divided in two broad categories: *quality in use*, consisting of 33 evaluation patterns, deals with technological and interaction characteristics of the system; *educational quality*, consisting of 36 evaluation patterns, refers to the degree to which a system supports effective teaching and learning. To give some examples, Table 2 presents the evaluation pattern QU_01 addressing quality in use, titled “Availability of communication tools”, while Table 3 presents the evaluation pattern EQ_19 addressing educational quality, titled “Quality of practical exercises”. In this framework, educational quality is defined by focussing on general best practices in the delivery of e-learning material rather than on specific knowledge of the domain, because the pattern has been defined to be as general as possible, independent of the specific topic of the e-learning system. Evaluation patterns specialized on a certain topic can be defined if considered necessary. For instance, EQ_19 drives the inspector’s attention on whether practical exercises are provided, if they use a consistent terminology, etc.

Table 2. An evaluation pattern addressing quality in use

Table 3. An evaluation pattern addressing educational quality

As we have discussed above, evaluation patterns are defined independently of design patterns. However, by comparing the proposed evaluation patterns with the pedagogical design patterns defined in the E-Len project (E-Len project, 2005), it can be noticed that the evaluation patterns actually address the problems mentioned in those design patterns and guide evaluators in verifying if the suggested solutions have been implemented.

4. Method

This section describes a study which compared the PB inspection (PB) against heuristic evaluation (HE) and thinking aloud (TA). A preliminary analysis of a small sub-set of the data was presented in (Ardito et al., 2006). This paper substantially expands our previous work proposing a comprehensive comparative metric, and providing an in depth analysis of the evaluators’ performance.

4.1 Participants

The study involved a total of 98 participants recruited from undergraduate students of the University of Bari in Italy. Specifically, 73 acted as novice evaluators, participating in the study as part of a course-work assignment for an advanced HCI course. They had basic knowledge of usability evaluation techniques, and previous experience evaluating web-sites using Nielsen’s heuristics (Nielsen and Tahir, 2002). They did not have any specific background on quality of e-learning systems. The remaining 25 participants were freshmen students, who acted as end users in the TA condition; these students did not have knowledge of usability or interaction design.

4.2 Design

Evaluation technique (3: PB, HE, and TA) was manipulated between-subjects. Evaluators were randomly assigned to a condition before the study took place. Two groups of 25 students each participated in the PB and TA conditions. The remaining 23 participated in the HE condition.

4.3 Procedure

A week before the study, the participants who acted as evaluators were given a 1-hour group demonstration of the application to be analysed. This addressed summary information about the application content and its main interactive functions. Two days before the study, a 1-hour training session introduced the evaluators to the specific technique they had to use during the evaluation. The study consisted of two sessions of three hours each: evaluation and output consolidation.

During the evaluation session, participants, tested in separate computer laboratories, were asked to evaluate the e-learning system by applying the technique they were assigned to. Data were collected in group settings, but every evaluator worked individually. We used four large laboratories (18m x 15m) of the Computer Science Department in Bari, since we needed two laboratories for the TA condition. Each laboratory provided 35 workstations on individual desks arranged in 7 rows and 5 columns, and a main desk with another workstation in the front. Participants in the TA group were introduced to the person who acted as their end user and invited to sit in one of two close laboratories. Each laboratory hosted 12 evaluator/user pairs, who were spread through the room leaving one empty row and one empty column between pairs. The 25th pair was accommodated at the main desk in one of the two laboratories. The pairs were far enough each other to avoid any interference. Each evaluator observed a student performing seven tasks, which were predefined in order to be equivalent in coverage to the evaluation patterns used in PB condition. The PB group, working in a third laboratory, received a list of eight evaluation patterns to be applied during inspection (Table 4). The limited number of evaluation patterns was due to time constraints. We selected evaluation patterns addressing to the analysis of the main features of the e-learning system. Finally, the HE group, working in the fourth laboratory, was provided with the ten "Learning with software" heuristics (Squires and Preece, 1999).

Table 4. The eight evaluation patterns tested in the study. The complete patterns are in (Lanzilotti et al., 2009)

Evaluators recorded problems on a booklet which differed according to the evaluation condition. The HE booklet included ten forms, one for each e-learning heuristic. The form required information about the interface location where the heuristic was violated and a short description of the problem. The PB group was provided with a booklet including eight forms, each one corresponding to an evaluation pattern. The forms required information about the violations detected through the specific evaluation pattern and where they occurred. The TA booklet included seven forms, one for each predefined task. The experimenter listed and described all problems that the user encountered performing a task. At the end of the evaluation session, all forms were collected.

During the output consolidation session, held the day after, all evaluators typed the content of the booklet in an electronic format to avoid readability problems during data analysis and standardise the output across the three conditions. For each problem, evaluators reported a description, where it occurred, and how it was found. They also ranked the problem severity from 1 (not severe at all) to 5 (very severe). Finally, participants filled in the evaluator-satisfaction questionnaire proposed in (De Angeli et al., 2003).

4.4 Application

Star Learning, a web-based e-learning system, was used as the target of the evaluation (StarLearning, 2009). The system provides access to on-line courses, auto-evaluation tests and exams. Logging into the system, students became part of a Virtual Classroom composed of all the students who were registered to the same course. The system offered several synchronous and asynchronous communication tools (chat, e-mail, forum) and tools for exercise and auto-evaluation. Participants in the experiment evaluated seven modules of a course on information technology, for a total of 100 pages.

4.5 Data coding

Two expert usability evaluators with a PhD in Human-Computer Interaction independently examined all the electronic booklets to identify individual and unique problems. They also scored each problem for severity on the same scale used by the participants, and classified them according to their cause and characteristics. The inter-rater reliability on each variable was satisfactory (>.80) and all differences were solved by discussion. One evaluator also performed a content analysis of problems based on their clarity and suggestions for design. Double-scoring was conducted on 20% of these data, yielding a value superior to .85.

4.6 Comparison metric

The comparison metric was defined along six main dimensions expanding traditional psychometrics literature (Graziano and Raulin, 2004) with the work of (Hartson et al., 2003; De Angeli et al., 2003; Blandford et al., 2008), who provided an in-depth description of different variables to facilitate the comparison of usability evaluation techniques beyond the simplistic problem-count comparison. Our proposal provides a synthetic metric which includes most of the variables discussed in previous work under 6 basic main dimensions. Each of them includes several variables, as summarised in Table 5. These variables were selected based on the specificity of our study but they could easily be increased to account for different contexts. Operational definitions of each variable are in the Results Section.

The first three dimensions covered traditional concepts in classical test theory, namely *reliability*, *validity*, and *effective range*. *Reliability* refers to consistency of measurement. Good techniques must give consistent results independently of who is performing the evaluation. *Validity* refers to the capability of a technique to measure what is intended to measure, i.e., detect real interaction problems and provide a proper estimation of their severity. These two dimensions have been widely discussed in the usability literature and there is little controversy with regard to their utility in facilitating comparisons across different techniques. *Effective range* refers to the sensitivity of a technique to measure the event of interest with the desired precision. In the usability domain it encompasses well established variables, such as that of thoroughness and effectiveness, but also more innovative variables such as that of scope, the indication of the different kinds of issues identified by different techniques (Blandford et al., 2008).

The three basic psychometric qualities were supplemented by specific dimensions addressing pragmatic aspects of usability evaluation, namely *cost*, *design impact*, and *perceived value*. *Cost* measures the resources needed to perform the evaluation in terms of time and number of evaluators. *Design impact* provides an estimation of the effects of the evaluation output on system improvement. This dimension encompasses aspects of *persuasive power* (the ability of an evaluator to persuade a designer to modify an interface as a result of the evaluation output) and *downstream utility* (the usefulness of the evaluation on informing redesign) described in (Hartson et al. 2003; Blandford et al., 2008; Cockton and Woolrych, 2009). As the system did not undergo any modification as a result of the evaluation, these two important variables could not be directly assessed in our study. Yet we tried to provide an indirect estimation of some of the aspects included in them, analysing the clarity of the report and the quality of possible design suggestions provided in it. Clarity of report is a pre-condition for persuasive power: designers need to understand the problem if they have to be persuaded to make any change. Clarity of report was addressed subjectively by expert evaluators who coded the data (see Section 4.5) and more objectively by looking at verbal variability within the report, under the assumption that linguistic standardisation is an important aspect of information sharing between different domains (usability experts and designers).

An innovative dimension proposed by our metric is *perceived value.* It considers the subjective assessment of the quality of a technique by the evaluators, in terms of satisfaction with it.

Table 5. Comparative metric

4.7 Hypotheses

The overarching hypothesis driving our research is that the PB inspection has the potential to improve the performance of novice evaluators by providing a systematic framework and clear indications on how to inspect an interactive system and report the evaluation results. Within this framework, we stated specific hypotheses relative to the performance of PB, HE and TA for each of the six quality dimensions addressed by the comparative metric. These hypotheses are based on previously reviewed literature and our own experience with e-learning evaluation studies. They are summarised in Table 6 and discussed below.

Table 6. Hypotheses

We expected an advantage of PB over the other evaluation techniques on 4 out of 6 evaluation dimensions. In particular, we assumed that the systematic nature of PB will impact on *reliability*, where we expected that PB will achieve the highest performance, followed in the order by TA and HE (*H1*). This order is due to the tendency of HE to find more problems of less serious nature than TA. We posited that the detection of less serious problems is more likely to be affected by subjective preferences (Ling and Salvendy, 2009) to the detriment of consistency. Following this line of reasoning, we also predicted that TA will be more *valid* than HE (*H2*) but we assumed that PB could counter-act the occurrences of false alarms by providing a structure and clear instructions to inspectors. A positive effect of PB on *design impact* was also hypothesised, as the activity prompts and the structured reporting format were specified in order to improve a common understanding between usability experts and designers, and to foster design-oriented thinking (*H3*). Finally, this general positive trend was deemed to have an effect on *perceived value*, as people applying evaluation patterns will feel more confident in their results (*H4*).

However, we expected two basic limitations of PB as compared to the other evaluation techniques. The major one refers to the *effective range,* which could be negatively affected by the limited number of evaluation patterns used in the study. We posited that evaluation patterns may have the undesirable effect to lead evaluators to focus only on selected aspects of the interface, disregarding other potentially important aspects not directly covered in the pattern (*H5*). Similarly, we anticipated that the *cost* of performing a PB inspection will be higher as compared to heuristic evaluation, but lower than thinking aloud (*H6*).

5. Results

A total of 217 unique problems and 38 non-problems (statements which reported not understandable content or unverifiable information) were found. On the average, these problems were scored as being of mixed severity (level 3) by the participants in the study and the experts who coded the data. The distribution of problems in the five severity categories based on the expert coding is reported in Table 7.

Table 7. Distribution of problems in the five severity categories

Detailed statistical analyses for each of the six dimensions of the comparison metrics are reported in this section. Section 6 discusses the results comparing them with the hypotheses indicated in Table 6.

5.1 Reliability

Reliability was measured with regard to *consistency of problems* found by the evaluators and to *consistency of severity rating*. The first variable was computed applying the any-two agreement formula (Hertzum et al., 2003)

$$\textit{Any-two agreement} = \text{Average of } \frac{|P_i \cap P_j|}{|P_i \cup P_j|} \text{ over all } 1/(2n(n\text{-}1)) \text{ pairs of evaluators}$$

where P_i and P_j are the set of problems detected by evaluator i and evaluator j, and n is the total number of evaluators. This measure ranges from 0%, if no two evaluators reported any problem in common, to 100% if all evaluators reported the same set of problems.

The any-two agreement index was analysed by an ANOVA with evaluation technique (3 levels) as between-subjects factor. The test returned a large significant effect for evaluation technique $F_{(2,70)} = 26.85$, $p < .001$, partial $\eta^2 = .43$. Descriptive statistics are summarised in Table 8. Post-hoc analysis, based on the LSD method, indicated a significant increase from HE to TA and from TA to PB, showing that patterns increased the reliability of the evaluation.

Table 8. Descriptive statistics of the any-two agreement index as a function of evaluation technique

Consistency of severity ratings was measured by the Intraclass Correlation Coefficient (ICC) for each experimental condition. This coefficient is used when a set of n targets are rated by k evaluators and indicates the correlation between one measurement on a target and another measurement obtained on that target (Shrout and Fleiss, 1979). The mathematical model used in the computation of the coefficient was based on a one-way analysis of variance (Case 1 analysis), which applies to cases where each target is rated by a different set of k judges, randomly selected from a larger population of judges. The ICC index ranges from a minimal of 0 (no agreement at all) to a maximum of 1 (perfect match). For each experimental condition, we randomly selected 10 problems which were rated by at least 5 evaluators (it is worth noting that only 1 problem was shared among the three analyses). Results indicated little (if any) correlation for HE and TA (*ICC* average measures = .10 and .17, respectively) and low correlation for PB (*ICC* average measures = .39).

To conclude, both analyses addressing reliability supported H1, showing that evaluators were more consistent in the problems they found and in the evaluation of their severity when using the PB inspection than when using other techniques (PB > TA > HE).

5.2 Validity

Validity was measured relative to the detection of usability problems (*P_Validity*) and their severity rating (*S_Validity*). The first variable was computed as

$$P\_Validity_i = \frac{P_i}{I_t}$$

where P_i is the number of real problems found by the i^{th} inspector, and I_t is the total number of issues identified as problems by that inspector. This variable presented a seriously skewed distribution as a very limited number of non-problems were collected. As a consequence, it was analysed by a Kruskal-Wallis H test, the nonparametric analogue of a one-way analysis of variance that can be applied to the comparison of 3 or more independent samples. The analysis returned a non significant effect of evaluation technique ($\chi^2 = 3.75$, $p = .15$), although the trend of results was in the expected direction (PB = TA > HE). The direct comparison between TA and HE by a Mann-Whitney U test returned a marginally significant effect $Z = -1.89$ (N = 48) $p = .06$ showing that TA tended to provide more valid results than HE. The same analysis contrasting TA and PB returned no significant results. This trend of results partially supported H2 (PB = TA > HE).

Validity of severity rating (*S_Validity*) was computed using an evaluation performed by experts as base-line criteria. The following formula was applied

$$S\_Validity = S_{ip} - S_{ep}$$

where S_{ip} is the severity rating of the i^{th} evaluator to a given problem *P* and S_{ep} is the severity rating of the expert coders to the same problem. Positive values denote underestimation and negative values overestimation. S_Validity was tested by an analysis of variance, with evaluation technique as between-subjects factor. A large significant effect of evaluation technique was found

$F_{(2,941)} = 5.96, p < .01$, partial $\eta^2 = .13$. LSD post-hoc analysis indicated a significant difference between PB (mean = .33, std error = .06) and both TA (mean = .11, std error = .08) and HE (mean = .02, std error = .08). These results are in contrast with H2, as they indicated that PB tended to induce overestimation of problem severity, whereas both TA and HE achieved more valid results.

5.3 Design impact

Three variables were taken into consideration to address design impact, namely *clarity of report*, *design suggestions*, and *linguistic variability*. The first two were based on a qualitative assessment by the two expert evaluators who analysed a random sample of 50 problems for each experimental condition. Problems were scored for clarity (on a three point scale: low, medium, high). Furthermore, the presence of explicit suggestion of design was noted (on a nominal scale *present*, *absent*). Finally, all reports were analysed by textpro (http://textpro.fbk.eu/), a suite of modular natural language processing tools for analysis of Italian and English texts (Pianta et al., 2008). Textpro returned the number of unique tokens contained in the reports and identified their grammatical functions (e.g., verbs, nouns, adjectives, prepositions, etc.). A measure of linguistic variability was obtained by counting the number of unique verbs, nouns, and adjectives and dividing it by the total number of verbs, nouns, and adjectives contained in the report:

$$Linguistic\_Variability = \frac{Unique\_Verbs + Unique\_Nouns + Unique\_Adjectives}{Total\_Verbs + Total\_Nouns + Total\_Adjectives}$$

Clarity of report was analysed by a Kruskal-Wallis H test, which showed a significant effect of evaluation technique ($\chi^2 = 7.24, p < .05$). This effect was mainly due to the highest clarity of the reports written by participants in the PB condition (mean rank = 84.32), followed in the order by HE (mean rank = 78.83) and TA (mean rank = 63.35). Only 10% of the problems contained some explicit design suggestions and these problems were detected by the three experimental conditions in a similar number.

The style of the report was also very different across experimental conditions. The average number of tokens used for describing a problem was significantly lower in the HE condition (mean = 20.82; std error = 2.21) than in the PB (mean = 31.30; std error = 2.12) and in the TA conditions (mean = 28.25; std error = 2.21), $F_{(2,70)} = 6.14, p < .05$, partial $\eta^2 = .15$. All grammar forms followed this trend of results, showing that participants in the PB and TA conditions were more verbose than participants in the HE conditions. Furthermore, participants in the PB condition showed significantly less linguistic variability (mean = 56%, std error = .016) than participants in the TA (mean = 61%, std error = .016) and HE conditions (mean = 64%, std error = .017), $F_{(2,70)} = 6.35$, $p < .01$, partial $\eta^2 = .15$. The highest homogeneity in the PB reports was due to the reuse of the standard terminology proposed in the evaluation patterns.

To conclude, the analyses on design impact supported H3, showing that the deliverables produced by evaluators exploiting patterns were clearer and more standardised than that produced by other evaluation techniques. However, no support for the hypothesis that patterns would foster design suggestions was found.

5.4 Perceived value

Perceived value was measured by a combination of *quantitative* and *qualitative data*. The quantitative measure 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 reflected the perceived *reliability* of the evaluation technique, the second factor the *gratification* derived by its use and the third factor its *ease 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 are illustrated in Figure 1.

Mean scores to the three factors were entered as dependent variables in a multivariate analysis of variance with evaluation 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 the three evaluation dimensions changed according to the technique. The univariate effects suggested that this change was principally due to the dimension of gratification ($F_{(2,72)} = 5.39$, $p < .01$). Post-hoc tests indicated that TA was evaluated as the most pleasant technique ($p < .05$) with no difference between the other two conditions (TA > PB = HE).

Figure 1. Average of reliability, gratification, and ease of use evaluations as a function of the evaluation technique

Qualitative data were collected from the final section of the questionnaire which invited evaluators to write down the best and the worst features of the technique they had used in the evaluation. A total of 78 positive and 72 negative comments were collected. A grounded analysis of these comments allowed the identification of 3 common themes used by participants to assess positive and negative aspects.

- *Perceived thoroughness* refers to the users' opinion on the coverage of the technique. Comments in this category addressed the extent to which a technique was perceived as being able or unable to highlight as many of the existing interaction problems of the e-learning system as possible.

- *Required expertise* refers to previous skills, background knowledge, and amount of training deemed necessary for applying the technique.
- *Gratification* regards emotional reactions to the activity.

Table 9 reports frequency (f) and percentage (%) values of positive (plus) and negative (minus) themes in the three experimental conditions (highest values within each condition are highlighted). References to perceived thoroughness and required expertise emerged in several comments addressing all the techniques, whereas the gratification theme emerged only in the TA condition where participants reported to have enjoyed the interaction with the user, as well as observing and trying to understand his/her behaviour. Some 78% of negative comments reported by participants in the TA condition referred to required expertise, as participants acknowledged that the outcome was strongly dependent on the evaluator skills and the user's characteristics. On the other hand, they clearly recognised the importance of involving real users in order to discover many interaction problems, as highlighted by many positive comments on thoroughness.

Participants using the PB inspection reported several positive comments (69%) related to required expertise, stressing the utility of evaluation patterns in guiding inspectors. In fact, only 5% of the comments addressing the expertise theme were negative. The strong majority of negative comments was related to thoroughness, as participants worried that the limited number of patterns used in the evaluation may have hampered the completeness of results. Thoroughness was the prevalent theme in the HE condition, both for positive and negative appraisal. Participants commented that heuristics were useful to discover problems of different types, and strongly appreciated the flexibility of the technique. However, they also acknowledged that heuristics were too general and underspecified, thus requiring some form of background knowledge.

Table 9. Frequency and percentage values of positive (Plus) and negative (Minus) comments reported for each category

To summarize, H4 was rejected, as the only subjective difference between the evaluation techniques indicated an advantage of thinking aloud over the other two techniques related to the gratification dimension.

5.5 Effective range

Effective range was addressed by four different variables: *thoroughness*, *serious thoroughness (S_Thoroughness), scope, and effectiveness*. Thoroughness refers to the completeness of the evaluation results with respect to the total number of real usability problems affecting the system (Hartson et al., 2003). This value was computed by the following formula

$$Thoroughness_i = \frac{P_i}{P_t}$$

where P_i is the number of problems found by the i^{th} inspector, and P_t is the total number of problems existing in the application (n = 217). S_Thoroughness refers to the completeness of the evaluation results with respect to high-severity problems. It was computed by the following formula, where *s* refers to severity ≥ 4.

$$S\_Thoroughness(s) = \frac{number\ of\ real\ problems\ found\ at\ severity\ level\ (s)}{number\ of\ real\ problems\ that\ exist\ at\ severity\ level\ (s)}$$

The thoroughness indexes were analysed by two separated analyses of variance with evaluation technique as between-subjects factor. In both cases, the effect of evaluation technique was significant (*Thoroughness* $F_{(2,70)} = 25.38, p < .001$, partial $\eta^2 = .42$; *S_Thoroughness(s)* $F_{(2,70)} = 4.21, p < .05$, partial $\eta^2 = .11$). Descriptive statistics of both variables as a function of experimental conditions are reported in Table 10. PB inspection consistently scored the highest mean values. LSD post-hoc comparisons revealed the following differences ($p < .05$): *Thoroughness* PB > TA = HE, and *S_Thoroughness(s)* PB = TA > HE.

Other interesting findings can be identified by inspecting Table 10. Firstly, overall the thoroughness of all the evaluation techniques was very low, probably due to the large number of pages to be inspected and the many problems present in the application. Secondly, only HE did not change value between the two indexes, confirming previous findings that HE tends to identify mostly low severity problems.

Table 10. Descriptive statistics for the thoroughness indexes as a function of evaluation technique

This positive trend of results in favour of PB is challenged however by a qualitative assessment of its performance with regard to very severe problems. This analysis showed the expected effect of attentional fixity which can be induced by the application of evaluation patterns. In the study, six problems were identified as usability catastrophes (5 in severity rating). They concerned: (1) the lack of mechanisms to signal the user position of navigation leading to disorientation; the inconsistent usage of (2) ambiguous icons and(3) field captions; (4) a system crash which regularly occurred whenever the user performed a specific command combination; a difficulty in interacting with some system functionalities, such as lack of text-editing functionalities when (5) performing an exercise or(6) writing a message in the forum. Problems 1, 2, and 3 are of primary importance in e-learning systems and were identified by 9 participants in the HE condition, 7 in the TA condition, and 5 in the PB inspection condition. Problems 4, 5, and 6 were identified by 5 participants in the TA condition, 2 in the HE condition but no participant in the PB condition. This mismatch may be due to the lack of specific evaluation patterns focusing on the problems.

In order to assess the scope, problems were divided into the two basic categories considered by the PB inspection: quality in use and educational quality. Quality in use covered technological and usability problems. Educational quality covered issues related to content, including statements related to quality of the information provided by the system in terms of subject-matter clarity and completeness, as well as information architecture. Table 11 reports frequency and percentage scores of problems in the two categories as a function of evaluation condition.

Table 11. Frequency and percentage of usability problems classified by category and evaluation condition

The total number of problems discovered by participants in the two categories (quality in use and educational quality) was entered as dependent variable in a mixed-design ANOVA with evaluation technique (3 levels being) as between-subjects factor and category (2 levels being) as within-subjects factor. The main effect of category was strongly significant $F_{(1,70)} = 232.32$, $p < .001$, partial $\eta^2 = .77$. All evaluators found much more problems related to quality in use than to educational quality. The main effect of technique was also strongly significant $F_{(2,70)} = 25.38$, $p < .001$, partial $\eta^2 = .42$. PB inspection consistently found more problems than the other techniques. Finally, there was a weak significant interaction between category and technique: $F_{(2,70)} = 1.61$, $p < .05$, partial $\eta^2 = .04$. This interaction is displayed in Figure 2, showing that it is due to TA and HE. These two techniques found the same number of quality in use problems, but TA found less educational quality problems than HE. PB inspection found more problems in all dimensions.

Figure 2. Average score of quality in use and educational quality problems as a function of evaluation technique

Problems related to quality in use were further divided into 4 categories according to their cause (Table 11): (a) graphical design (adverse comments on aesthetic aspects of the interface); (b) feedback (negative statements addressing communication between the user and the interface); (c) navigation (problems related to the appropriateness of mechanisms for accessing information and for getting oriented in the system) and; (d) technology issues (e.g. page visualization, compatibility of the system with the browser, downloading time).

The frequency of problems in the three usability related dimension (graphical design, feedback and navigation) were analysed by a mixed-design ANOVA with evaluation technique (3 levels being) as between-subjects factor and category (2 levels being) as within-subjects factor. The main effect of technique was significant $F_{(1,70)} = 9.00, p < .001$, partial $\eta^2 = .20$, confirming that PB discovered more usability problems than the other techniques with no difference between them. The 2-way interaction was strongly significant $F_{(4,70)} = 16.17, p < .001$, partial $\eta^2 = .37$. Mean values are illustrated in Figure 3, supporting the idea that different techniques tend to identify different usability problems.

Figure 3. Average score of quality in use problems as a function of evaluation technique

The effectiveness variable captured the simultaneous effect of thoroughness and validity. It was defined as the product of thoroughness and validity as reported in (Hartson et al., 2003):

Effectiveness = Thoroughness x *Validity*

The ANOVA returned a strong significant effect for evaluation technique ($F_{(2,70)} = 24.29, p < .001$, partial $\eta^2 = .41$), due to the score of the PB (mean = .08, std error = .005) being significantly different from both HE (mean = .05, std error = .005) and TA (mean = .04, std error = .005) (PB > HE ≥ TA).

To conclude, H5 was rejected due to the unexpected tendency of PB to highlight a large number of problems of different type. However, some evidence of attentional fixity still emerged from the analysis of the most serious problems.

5.6 Cost

The cost dimension included two efficiency measures, one dealing with *evaluation time* and the other one dealing with the minimal *number of evaluators* who would enable the detection of a reasonable percentage of problems in the application.

Evaluation time was measured considering the average number of problems each participant found in 10 minutes (the whole evaluation lasted 180 minutes). The ANOVA indicated that overall this measure is affected by experimental condition, $F_{(2,70)} = 3.80, p < .05$. Post Hoc comparisons (LSD) showed that PB was the most efficient technique (mean = 1.19 problems in 10 minutes, std error = .08). There were no differences between HE and TA, mean = .91, std error = .08 and mean = .90, std error = .08, respectively.

The minimal number of evaluators was analyzed by plotting the cost-benefit curve proposed by Nielsen and Landauer (1993). It relates the proportion of usability problems to the number of evaluators applying the following formula

$Found(i) = n(1-(1-\lambda)^i)$

where *Found(i)* is the number of problems found by aggregating reports from *i* independent evaluators, *n* is the total number of problems in the application, and λ is the probability of finding the average usability problem when using a single average evaluator. The cost-benefit curves for the three groups are reported in Figure 4 ($n = 217$, $\lambda_{HE} = 0.05$, $\lambda_{TA} = 0.04$, $\lambda_{PB} = 0.08$). It emerges that PB consistently reached a better performance with the lowest number of evaluators, while HE and TA were more similar in performance. Assuming the standard 75% threshold considered to indicate maximum efficiency (Nielsen and Landauer, 1993), it emerged that PB reached it with 15 evaluators, whereas neither HE nor TA reached it with more than 23 evaluators.

Thus, H6 was rejected. Indeed, despite our hypothesis that a rigorous application of evaluation patterns is costly, participants in the PB condition performed better than expected.

Figure 4. The cost-benefit curve for the three evaluation techniques

6. Discussion

The objective of this study was to assess the value of evaluation patterns for facilitating the evaluation of e-learning systems. For this purpose, we have analysed a large sample of inexpert evaluators asking them to evaluate an e-learning application using three different techniques: Pattern-Based inspection, heuristic evaluation and thinking aloud. The comparison metric was defined along six different dimensions that extended traditional psychometric properties with usability specific attributes. Results suggested that evaluation patterns have the potential to improve the evaluators' performance as compared to the other two evaluation techniques, in terms of reliability, design impact, effective range, and cost. Furthermore, they increased validity in terms of problems discovered but not of severity rating. A summary of the results, with reference to the original hypotheses, is reported in Table 12.

Table 12. Results of the comparison study

The first hypothesis was fully confirmed. The PB inspection was the most reliable technique, followed in the order by thinking aloud and heuristic evaluation. Participants who exploited evaluation patterns obtained more homogeneous results both in terms of types of problems and severity estimation. This result is in contrast with other studies (see Table 1) which revealed an advantage of usability testing over the analytical approach in terms of problems severity (Steve et al., 2001; Hornbæk and Frøkjær, 2005; Tan et al., 2009), demonstrating that evaluation patterns provide a robust framework to inspectors applying analytical approach. It has to be noted however that the average reliability of the evaluation techniques was very low (36% in the best condition) suggesting that the evaluation was strongly affected by individual variations. This is not a new finding in the literature where values as low as 5% are reported (Hertzum et al., 2003). We believe that, in our study, limited reliability was due to the large application to be inspected in a short amount of time, and to the little expertise of the evaluators.

The hypothesis concerning validity was only partially supported. As expected, the PB inspection and thinking aloud were found to be equal and better than heuristics evaluation, but only with regard to problem detection. On the contrary, the analysis of the validity of severity estimation showed that PB inspection tended to overestimate problems, while thinking aloud and heuristic evaluation obtained more valid results. This was an unexpected result, which may be due to the different types of problems evinced in the three conditions, or to a judgment bias induced by the strict methodological apparatus available to inspectors using evaluation patterns.

The hypothesis relative to design impact was partially supported. Participants in the PB condition produced clearer and more standardised reports, as compared to the other two techniques. Thinking aloud was the worst condition with regard to output readability. The positive result of the PB inspection can be attributed to the item *Output* of the evaluation patterns, which suggests the terminology for reporting the inspection results. While in the case of heuristic evaluation, and more importantly of thinking aloud, evaluators had little to no guidance. However, contrary to our expectations, no differences emerged with regard to the provision of design suggestions, which were extremely rare in all conditions.

The hypothesis relative to perceived value was rejected. The questionnaire analysis showed an advantage of thinking aloud on the gratification dimension due to the social nature of the evaluation setting. The qualitative analysis of participants' comments revealed three major themes used to talk about evaluation techniques: thoroughness, required expertise and gratification. Participants applying the PB inspection expressed their satisfaction to be able to carry out a good evaluation because patterns guided them during their work. At the same time, however, they worried about the evaluation not covering all problems due to the limited number of patterns. This is an obvious limitation, at the basis of the hypothesis on effective range (H5), which, however, received only partial empirical support in the study. Consistently with the questionnaire results, thinking aloud was found to be the most gratifying technique, but participants worried about the dependency of the results on the evaluator expertise and the user characteristics. Finally, participants in the heuristic condition declared that the general formulation of the heuristic was both a positive and a negative. Indeed if, on the one hand, it permits to discover problems of different nature, on the other hand, it requires a certain level of expertise to be correctly applied.

The two negative hypotheses relative to effective range and costs were rejected, as PB inspection performed better than expected. Indeed, with regard to effective range, it was found to be the best technique in terms of thoroughness, serious thoroughness, coverage of different types of problems and effectiveness. However, we found evidence of the expected fixation effect which may be induced by evaluation patterns, since most participants in this condition missed the most serious problems as they were not covered by the available patterns. The study also confirmed that heuristic evaluation is inclined to collect low severity problems. With regard to cost, the PB inspection resulted in the most efficient technique both in terms of time and number of evaluators necessary for discovering a reasonable number of problems. Despite a rigorous application of several evaluation patterns is time demanding, the highest efficiency of the PB inspection was due to the higher number of discovered problems.

7. Conclusions

The title of this paper poses the question whether patterns could help novice evaluators. The results of a study that compared the Pattern-Based inspection against two well known evaluation techniques (heuristic evaluation and thinking aloud) provide a promising answer to this question showing that patterns can indeed improve evaluation on a number of measurement qualities, including *reliability*, *validity* (at least partially), *effective range*, *design impact* and *cost.* The study suggests that patterns have the potential to reduce one of the main drawbacks of other inspection methods, namely their dependency on the evaluator's skills and experience. Patterns help share and transfer the evaluation know-how of expert inspectors, thus simplifying the inspection process for newcomers. They indicate how best to conduct an inspection, showing which aspects of the application the evaluators should concentrate on and prescribing operational activities. The study also pointed out the main disadvantages of patterns, such as the risk of disregarding even major usability problems because the evaluator attention is guided towards those aspects of the application directly addressed by the set of evaluation patterns used for the inspection. This risk can be limited by the application of a larger set of evaluation patterns.

The PB inspection is a technique applicable to the evaluation of any interactive systems, provided that a set of valid evaluation patterns addressing critical aspects of the application are provided. In this paper, the PB inspection has been used to evaluate e-learning system. To this aim, beside patterns addressing quality in use, also patterns related to the educational quality of such systems have been defined to support evaluators in analysing such aspects, which tend to be disregarded in user testing, as well as in other analytical methods, even when applying e-learning specific heuristics (Squires and Preece, 1999).

This paper also contributes a comprehensive metric for comparative studies of evaluation techniques which extend the psychometric literature with specific dimensions and may help standardisation of studies. Moreover, it fosters the discussion about the objectivity of usability assessment, brought forward by the UX research agenda (Hassenzahl and Tractinsky, 2006) which explicitly recognises the subjective nature of experiences, claiming that quality appraisal is modulated by a number of individual and contextual factors. This standpoint also recognises that the relationship between usability and other dimensions of the UX is complex and that judgements on one dimension can sometimes colour judgements on other dimensions consistently with what in psychology is known as the halo effect (De Angeli et al., 2006, De Angeli et al., 2009).

The study reported in this paper has some limitations which need to be taken into account when analysing the results. An obvious limitation regards the nature of the sample and the evaluation context. More research is needed to understand how these findings extend to real work-related settings. Furthermore, due to experimental constraints, our sample had to evaluate a very large application in a limited amount of time. Yet, this study reveals important findings supported by a large sample of users and proposes a strong methodological apparatus for future research.

Acknoledgement

The research reported in this paper has been supported by the following grants: COST Action 294 MAUSE (Towards the MAturation of Information Technology USability Evaluation), COST Action IC 0904 TWINTIDE (Towards the Integration of Transectorial IT Design and Evaluation), the Italian MIUR project L4ALL (Learning For All), and the Regional project DIPIS (DIstributed Production as Innovative System). We are grateful to the anonymous reviewers for their useful comments, which helped improving the quality of the presentation.

References

Alexander, C., Ishikawa, S., Silverstein, M., Jackobson, M., Fiksdhal-King, I., Angel, S., 1977. A pattern language. Oxford University Press, New York, USA.

Ardito, C., Costabile, M.F., De Angeli, A, Lanzilotti, R., 2006. Systematic evaluation of e-learning systems: an experimental validation. In: Proceedings of NordiCHI, Oslo, Norway, October 14-18. ACM Press, pp. 195-202.

Avgeriou, P., Papasalouros, A., Retalis, S., Skordalakis, M. 2003. Towards a pattern language for learning management systems. Educational Technology & Society 6 (2), 11-24.

Basili, V., Caldiera, G., Rombach, D. 1994. The goal question metric approach. Encyclopedia of software engineering. Wiley&Sons Inc., USA.

Bernstein, M., 1998. Patterns of hypertext. In: Proceedings of Hypertext and Hypermedia, Pittsburgh, USA, June 20-24. ACM Press, pp. 21-29.

Blandford, A., Hyde, J.K., Green, T.R.G., Connell, I. 2008. Scoping analytical usability evaluation methods: a case study. Human Computer Interaction Journal 23 (3), 278-327.

Borchers, J., 2001. A pattern approach to interaction design. John Wiley & Sons Inc., New York, USA.

Cockton, G., Woolrych, A. 2009. Comparing usability evaluation methods: strategies and implementation. In: Law, E., Scapin, D., Cockton, G., Springett, M., Stary, C., Winckler, M. (Eds). Maturation of usability evaluation methods: retrospect and prospect - Final Reports of COST294-MAUSE, pp. 18-82.

Costabile, M.F., Roselli, T., Lanzilotti, R., Ardito, C., Rossano, V. 2007. A holistic approach to the evaluation of e-learning systems. In: C. Stephanidis (Ed.), Universal Access in HCI, Part III. Vol. LNCS 4556, Springer-Verlag, Berlin Heidelberg, Germany, pp. 530-538.

De Angeli, A., Hartmann, J., Sutcliffe, A., 2009. The effect of brand on the evaluation of websites. In: Proceedings of Interact 2009, Uppsala, Sweden, August 24-28, pp. 638-651.

De Angeli, A., Matera, M., Costabile, M.F., Garzotto, F., Paolini, P., 2003. On the advantages of a systematic inspection for evaluating hypermedia usability. International Journal Human-Computer Interaction 15 (3), 315-335.

De Angeli, A., Sutcliffe, A., Hartmann, J., 2006. Interaction, usability and aesthetics: what influences users' preferences?. In: Proceedings of Designing Interactive Systems 2006, State College, Pennsylvania, USA, June 24-26. ACM Press, pp. 271-280.

Delta checklist, 2002. Available at: http://www-interact.eng.cam.ac.uk/CAL95/Eval-Checklist1.html. Last access: May, 2009.

Dimitriadis, Y., Goodyear, P., Retalis S. Editors. 2009. Design patterns for augmenting e-learning experiences. Special issue of Computers in human behaviour 25 (5), 997-1188.

Dix, A., Finlay, J., Abowd, G., Beale, R., 2003. Human-Computer Interaction, third ed. Pearson-Prentice Hall, Essex, UK.

Doubleday, A., Ryan, M., Springett, M., Sutcliffe, A., 1997. A comparison of usability techniques for evaluating design. In: Proceedings of Designing Interactive Systems, Amsterdam, The Netherlands, August 18-20. Springer Verlag, pp. 101-110.

E-Len Project, 2005. E-LEN: A network of e-learning centres. Available at: http://www2.tisip.no/E-LEN/. Last access: April, 2010.

Frøkjær, E. Hornbæk, K., 2008. Metaphors of human thinking for usability inspection and design. ACM Transaction on Computer-Human Interaction 14 (4), 1-33.

Gamma, E., Helm, R., Johnson, R., Vlissedes, J., 1995. Design patterns – elements of reusable object oriented software. Addison Wesley, Reading, USA.

Garzotto, F., Mainetti, L., Paolini, P., 1994. Adding multimedia collections to the dexter model. In: Proceedings of Hypertext and Hypermedia, Edinburgh, United Kingdom, September 19-23. ACM Press, pp. 70-80.

Gerdt, P., Miraftabi, R., Tukiainen, M., 2002. Evaluating educational software environments. In: Proceedings of the International Conference on Computers in Education, Auckland, New Zealand, December 3-6, pp. 675-676.

Gray, W.D., Salzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction 13 (3), 203–261.

Graziano, A.M., Raulin, M.L., 2004. Research methods – A process of inquiry. Pearson Education Group, New York, USA.

Hartson, H.R., Andre, T.S., Williges, R.C., 2003. Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction 15 (1), 145-181.

Hassenzahl, M., Tractinsky, N., 2006. User experience - a research agenda. Behaviour and Information Technology 25 (2), 91-97.

Hertzum, M., Jacobsen, N.E., 2003. The evaluator effect: a chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction 15 (1), 183-204.

Hornbæk, K., 2006. Current practice in measuring usability: challenges to usability studies and research. International Journal of Human-Computer Studies 64, 79-102.

Hornbæk, K., Frøkjær, E., 2005. Comparing usability problems and redesign proposals as input to practical systems development. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, Portland, Oregon, USA, April 02-07. ACM Press, pp. 391- 400.

Jeffries, R., Desurvire, H.W., 1992. Usability testing vs heuristic evaluation: was there a context? ACM SIGCHI Bulletin 24 (4), 39-41.

Jeffries, R., Miller, J. R., Wharton, C., Uyeda, K., 1991. User interface evaluation in the real world: a comparison of four techniques. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, New Orleans, Louisiana, USA, April 27-May 02. ACM Press, pp. 119-124.

Juristo, N., Moreno, A., Sanchez-Segura, M. 2007. Guidelines for Eliciting Usability Functionalities. IEEE Transactions on Software Engineering 33 (11), 744-758.

Kantner, L. Rosenbaum, S. 1997. Usability studies of WWW sites: heuristic evaluation vs. laboratory testing. In: Proceedings of the 15th Annual international Conference on Computer Documentation, Salt Lake City, Utah, USA, October 19-22. ACM Press, pp. 153-160.

Karat, C., Campbell, R., Fiegel, T., 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, Monterey, California, USA, May 03-07. ACM Press, pp. 397-404.

Lanzilotti, R., 2006. A holistic approach to designing and evaluating e-learning systems quality: usability and educational effectiveness. PhD Dissertation, Department of Computer Science, University of Bari, Italy.

Lanzilotti, R., Ardito, C., Costabile, M.F., De Angeli, A., 2009. The PB inspection: a comparison study. IVU Technical Report, September 2009, Department of Computer Science, University of Bari, Italy.

Lanzilotti, R., Costabile, M.F., Ardito, C., De Angeli A., 2006. eLSE methodology: a systematic approach to the e-learning systems evaluation. Educational Technology and Society 9 (4), 42-53.

Law, E. 2007. Heuristic Evaluation. In: Proceedings of COST294-MAUSE International workshop "Review, Report and Refine Usability Evaluation Methods (R3-UEM)", Athens, Greece, March 5, pp. 61-63.

Law, L., Hvannberg, E.T., 2002. Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. In: Proceedings of NordiCHI 2002, Aarhus, Denmark, October 19-23. ACM Press, pp. 71-80.

Law, E., Scapin, D., Cockton, G., Springett, M., Stary, C., Winckler, M. Editors. 2009. Maturation of usability evaluation methods: retrospect and prospect - Final Reports of COST294-MAUSE Working Groups.

Learning Technology Dissemination Initiative: Implementing Learning Technology, 2002. Available at: http://www.ltdi.hw.ac.uk/ltdi/implementing-it/. Last access: May 2009.

Liljegren, E., 2006. Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. International Journal of Industrial Ergonomics 36 (4), 345–352.

Lim, K.H., Bembasat, I., Tood, P.A., 1996. An experimental investigation of the interactive effects of interactive style, instructions, and task familiarity on user performance. ACM Transaction on Computer-Human Interaction 3 (1), 1-37.

Ling, C., Salvendy, G., (2009). Effect of evaluators' cognitive style on heuristic evaluation: field dependent and field independent evaluators. International Journal of Human-Computer Studies 67, 382-393.

Matera, M., Costabile, M.F., Garzotto, F., Paolini, P., 2002. SUE Inspection: an effective method for systematic usability evaluation of hypermedia. IEEE Trans. Systems, Man and Cybernetics - Part A 32 (1), 93-103.

Meira, L., Peres, F., 2004. A dialogue-based approach for evaluating educational software. Interacting with computers 16, 615-633.

Mendes, E., Hall, W., Harrison, R. 1998. Applying metrics to the evaluation of educational hypermedia application. Journal of Universal Computer Science 4 (4), 382-403.

Nanard, M., Nanard, J., Kahn, P., 1998. Pushing reuse in hypermedia design: golden rules, design patterns and constructive templates. In: Proceedings of the Conference on Hypertext and Hypermedia, Pittsburgh, Pennsylvania, USA, June 20-24. ACM Press, pp. 11-20.

Ng, V., Chan, S., Lee, J., So, K. 1999. Some useful metrics on evaluating educational hypermedia designs. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Tokyo, Japan, October 12-15, pp. 247-252.

Nielsen, J., 1993. Usability engineering. Academic Press, Cambridge, USA.

Nielsen, J., Landauer, T. K., 1993. A mathematical model of the finding of usability problems. In: Proceedings of INTERCHI'93, Amsterdam, the Netherlands, April 24-29. ACM Press, pp. 296-213.

Nielsen, J., Mack, R.L., 1994. Usability inspection methods. John Wiley and sons, New York, USA.

Nielsen, J., Phillips, V. L., 1993. Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In: Proceedings of INTERACT '93 and CHI '93, Amsterdam, The Netherlands, April 24-29. ACM Press, pp. 214-221.

Nielsen, J., Tahir, M. 2002. Homepage Usability: 50 Websites Deconstructed. New Riders Publishing, USA.

Notess, M., 2001. Usability, user experience, and learner experience. eLearn 2001 (8), 3.

Parlangeli, O., Marchigiani, E., Bagnara, S., 1999. Multimedia system in distance education: effects on usability. Interacting with Computers 12, 37-49.

Pascoe, J., Ryan, N., Morse, D., 2000. Using while moving: HCI issues in fieldwork environments. ACM Transaction on Computer-Human Interaction 7 (3), 417-437.

Pianta, E., Girardi C., Zanoli, R., 2008. The TextPro tool suite. In: Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference, 28-30 May 2008, Marrakech, Marocco.

Preece, J., Rogers, Y., Sharp, H., 2007. Interaction design, second ed. John Wiley & Sons, New York, USA.

Quinn, C.N., Alem, L., Eklund, J.A., 1997. Pragmatic evaluation methodology for an assessment of learning effectiveness in instructional systems. In: S. Bewster, A. Cawsey & G. Cockton (Eds.) Human-Computer Interaction, Vol. II. British Computer Society, Edinburgh Press, pp. 55-56.

Ravden, S., Johnson, G., 1989. Evaluating usability of human-computer interfaces, a Practical method. Ellis Horwood Limited, Chichester, UK.

Schwier, R.A., Misanchunk, E.R., 1993. Interactive multimedia instruction. Educational Technology Publications, Englewood Cliffs, USA.

Shrout, P.E. Fleiss, J.L., 1979. Intra-class correlations: uses in assessing rater reliability. Psychological bulletin 2, 420-428.

Squires, D., Preece, J., 1999. Predicting quality in educational software: evaluating for learning, usability, and the synergy between them. Interacting with Computers 11 (5), 467-483.

Ssemugabi, S., de Villiers, R., 2007. A comparative study of two usability evaluation methods using a web-based e-learning application. In: Proceedings of SAICSIT 2007, Port Elizabeth, South Africa. ACM Press, pp. 132-142.

StarLearning (2009). http://multimedialab.di.uniba.it/generale/login.asp. Last access: May, 2009.

Steves, M. P., Morse, E., Gutwin, C., Greenberg, S, 2001. A comparison of usage evaluation and inspection methods for assessing groupware usability. In: Proceedings of ACM SIGGROUP Conference on Supporting Group Work, Boulder, Colorado, USA, September 30 - October 03. ACM Press, pp. 125-134.

Tan, W., Dahai L., Bishu R., 2009. Web evaluation: heuristic evaluation vs. usability testing. International Journal of Industrial Ergonomics 39 (4), 621-627.

Tidwell, J. 1999. Common ground: a pattern language for human-computer interface design. Available at http://www.mit.edu/~jtidwell/common_ground.html. Last access: May, 2009.

Tidwell, J., 2005. Designing Interfaces: Patterns for Effective Interaction Design. O'Reilly Med. Incorporation.

Wixon, D., 2003. Evaluating usability methods: Why the current literature fails the practitioner. Interactions 10 (4), 29-34.

Wong, B., Nguyen, T.T., Chang, E., Jayaratna, N., 2003. Usability metrics for e-Learning. In: Proceedings of the Workshop on Human Computer Interface for Semantic Web and Web Applications, Catania, Sicily, Italy, November 3-7. Springer-Verlag, pp. 235-252..

Reference	*Method*	*Design*	*Participants*	*Dependent Variables*	*Results*
Jeffries et al. (1991)	HE	Between-subjects	4 evaluators	Number of problems	HE > DO > CW = SG
	DO		6 users	Severity rating	HE > DO > CW = SG
	CW		1 human factor experts	Time	HE > DO > CW = SG
	SG		3 usability engineers		
Karat et al. (1992)	TA	Between-subjects	24 evaluators	Number of problems	TA > TW > IW
	IW		24 users	Severity rating	TA > IW, TW
	TW		2 usability engineers	Time	TA < IW, TW
Nielsen and Phillips (1993)	HE	Between-subjects	25 evaluators	Number of problems	DO > HE = GOMS Hot HE > Cold HE
	DO		20 users		
	GOMS		19 students	Time	DO > HE, GOMS Hot HE > DO
Doubleday et al. (1997)	HE	Between-subjects	5 evaluators	Number of problems	HE > TA
	TA		20 users	Severity rating	TA > HE
				Time	TA > HE
Steves et al. (2001)	HE	Between-subjects	4 evaluators	Number of problems	HE = UT
	UT		5 users	Severity rating	HE = UT
				Time	UT > HE
Law and Hvannberg, (2002)	HE	Between-subjects	2 evaluators	Number of problems	HE > TA
	TA		10 users	Severity rating	TA > HE
Hornbæk and Frøkjær (2005)	MOT	Between-subjects	43 evaluators	Number of problems	MOT = TA
	TA			Severity rating	MOT < TA
Liljegren, E., 2006	HE	Between-subjects	\	Number of problems	TA > CW, HE, HTA CW = HTA
	TA				
	CW			Severity rating	TA > CW, HE, HTA
	HTA				
Frøkjær and Hornbæk (2008)	MOT	Within-subjects	58 evaluators	Number of problems	MOT = HE > CW, TA
	TA			Severity rating	MOT = CW > HE = TA
	CW				
Tan et al. (2009)	HE	Between-subjects	36 evaluators	Number of problems	HE = DO
	DO		48 users	Severity rating	HE = DO

Legend: CW: Cognitive Walkthrough, DO: Direct Observation, GOMS: Goals, Operators, Methods, and Selection, HE: Heuristic Evaluation, HTA: Hierarchical Task Analysis, IW: Individual Walkthrough, MOT: Metaphors Of human Thinking, SG: Software Guidelines, TA: Think Aloud, TW: Team Walkthrough, UT: User Testing.

Table 1. Summary of studies comparing analytical evaluation with usability testing, ordered by publication date

QU_01: AVAILABILITY OF COMMUNICATION TOOLS

Focus of action: communication tools

Intent: check if the e-learning system provides tools that permit communication among learners, teachers, tutors, etc., and verify their usability

Activity prompts:

- Navigating in the system, identify the available synchronous/asynchronous communication tools
- Being a learner, try to communicate with others (learners, teachers,..)
- Being a teacher, try to communicate with others (learners, teachers,..)
- Being a tutor, try to communicate with others (learners, teachers,…)

Output: a description reporting:

- If synchronous/asynchronous communication tools are not present
- Difficulties in identifying communication tools
- Difficulties to communicate with learners
- Difficulties to communicate with teachers
- Difficulties to communicate with tutors
- Inconsistencies among communication tools

Table 2. An evaluation pattern addressing quality in use

EQ_19: QUALITY OF PRACTICAL EXERCISES

Focus of action: practical exercises

Intent: verify the quality of exercises to allow students to practice with the learned content

Activity prompts:

- Open a course module
- Verify if practical exercises are provided
- Verify if exercises are adequate to the module content (e.g. consistency of topics, consistency of terminology, etc.)
- Perform exercises focusing on the provided feedback (e.g., results about the executed exercise, suggestions on errors, etc.)

Output: a description reporting:

- If exercises are not provided
- Inconsistencies between exercises and module content
- Problems about inappropriate feedback

Table 3. An evaluation pattern addressing educational quality

Code	***Title***
QU_01	Availability of communication tools
QU_02	Quality of graphical interface elements
QU_27	Availability of course evaluation tools
EQ_06	Course organization
EQ_19	Quality of practical exercises
EQ_24	Topic prerequisites
EQ_28	Feedback of evaluation tools
EQ_35	Quality of evaluation tools results

Table 4. The eight evaluation patterns tested in the study. The complete patterns are in (Lanzilotti et al., 2009)

Dimensions	***Variables***
Reliability	Consistency of problems Consistency of severity rating
Validity	Validity of problems Validity of severity rating
Effective range	Thoroughness

	Serious thoroughness Scope Effectiveness
Cost	Evaluation time Number of evaluators
Design impact	Clarity of report Design suggestions Linguistic variability
Perceived value	Evaluator satisfaction quantitative data Evaluator satisfaction qualitative data

Table 5. Comparative metric

Hypothesis number	*Dimensions*	*Proposition*
H1	Reliability	PB > TA > HE
H2	Validity	PB = TA > HE
H3	Design impact	PB > HE > TA
H4	Perceived value	PB > TA > HE
H5	Effective range	TA > HE > PB
H6	Cost	TA > PB > HE

Table 6. Hypotheses

Severity 1 *Not serious at all*	*Severity 2*	*Severity 3*	*Severity 4*	*Severity 5* *Very severe*
17	53	79	62	6

Table 7. Distribution of problems in the five severity categories

Evaluation technique	*Mean*	*Std. Error*
HE	20.35	1.58
TA	27.48	1.52
PB	36.35	1.52

Table 8. Descriptive statistics of the any-two agreement index as a function of evaluation technique

Theme	***Heuristic Evaluation***				***Thinking Aloud***				***PB inspection***			
	Plus		**Minus**		**Plus**		**Minus**		**Plus**		**Minus**	
	f	%	f	%	f	%	f	%	f	%	f	%
Perceived thoroughness	13	52	14	67	13	48	4	15	8	31	21	95
Required expertise	12	48	6	29	3	11	21	78	18	69	1	5
Gratification	0	0	1	5	11	41	2	7	0	0	0	0
	25	100	21	100	27	100	27	100	26	100	22	100

Table 9. Frequency and percentage values of positive (Plus) and negative (Minus) comments reported for each category

Evaluation Technique	*Thoroughness*		*S_Thoroughness(s)*	
	Mean	Std. Error	Mean	Std. Error
HE	.051	.005	.051	.015
TA	.040	.005	.080	.015
PB	.086	.005	.113	.015

Table 10. Descriptive statistics for the thoroughness indexes as a function of evaluation technique

Category		*Heuristic Evaluation*		*Thinking Aloud*		*PB inspection*	
		f	%	f	%	f	%
Quality in use	Total	89	85	67	81	102	73
	Graphical design	22	21	15	18	40	29
	Feedback	43	41	27	33	32	23
	Navigation	18	17	15	18	16	11
	Technology	6	6	10	12	14	10
Educational quality	Total	16	15	16	19	38	27
	Total problems	105	100	83	100	140	100

Table 11. Frequency and percentage of usability problems classified by category and evaluation condition

Hypothesis number	*Dimension*	*Proposition*	*Results*		
H1	Reliability	PB > TA > HE	Supported		
H2	Validity	PB = TA > HE	Partially supported	Validity of problems	PB = TA > HE
				Validity of severity rating	TA = HE > PB
H3	Design impact	PB > HE > TA	Partially supported	Clarity of report	PB > HE > TA
				Design suggestions	PB = HE = TA
				Linguistic variability	PB > HE > TA
H4	Perceived value	PB > TA > HE	Not supported		TA > PB = HE
H5	Effective range	TA > HE > PB	Not supported	Thoroughness	PB > HE = TA
				Serious thoroughness	PB = TA > HE
				Effectiveness	PB > HE ≥ TA
				(Evidence of attentional fixity affecting PB)	
H6	Cost	TA >PB > HE	Not supported		PB > HE = TA

Table 12. Results of the comparison study