Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 19 (2007) 225–240
Interacting with Computers
www.elsevier.com/locate/intcom

Heuristic evaluation: Comparing ways of finding and reporting usability problems

Ebba Thora Hvannberg [a,*], Effie Lai-Chong Law [b], Marta Kristín Lárusdóttir [c]

[a] University of Iceland, Hjardarhaga 2-6, 107 Reykjavik, Iceland
[b] Eidgenössische Technische Hochschule Zürich Gloriastrasse 35, CH-8902, Zürich, Switzerland
[c] Reykjavik University, Ofanleiti 2, 103 Reykjavik, Iceland

Received 28 June 2005; received in revised form 29 August 2006; accepted 11 October 2006
Available online 1 December 2006

Abstract

Research on heuristic evaluation in recent years has focused on improving its effectiveness and efficiency with respect to user testing. The aim of this paper is to refine a research agenda for comparing and contrasting evaluation methods. To reach this goal, a framework is presented to evaluate the effectiveness of different types of support for structured usability problem reporting. This paper reports on an empirical study of this framework that compares two sets of heuristics, Nielsen's heuristics and the cognitive principles of Gerhardt-Powals, and two media of reporting a usability problem, i.e. either using a web tool or paper. The study found that there were no significant differences between any of the four groups in effectiveness, efficiency and inter-evaluator reliability. A more significant contribution of this research is that the framework used for the experiments proved successful and should be reusable by other researchers because of its thorough structure.

Keywords: User interface; Heuristic evaluation; Reporting; Web tool; Effectiveness; Efficiency; Comparison framework

1. Introduction

Since the early 1990s, researchers have carried out studies comparing and contrasting some of the methods brought forward to uncover usability problems of interactive computer systems, (Desurvire et al., 1992; Holzinger, 2005; Jeffries et al., 1991; Karat et al., 1992). Current research on usability evaluation clearly searches for methods that produce beneficial results for users and developers alike at low cost in an ever-increasing competitive industry.

In this paper, we report a case study of a framework for validating the use of usability evaluation methods and of problem registration tools and other support for enabling structured usability problem reporting. The focus is on refining a research agenda for comparing and contrasting evaluation methods. In the case study, the number and the seriousness of problems found per evaluator in heuristic evaluation, with two different sets of usability heuristics: Nielsen's heuristics and the cognitive principles of Gerhardt-Powals (Gerhardt-Powals, 1996), were compared. Furthermore, two different ways of reporting usability problems, on paper and with the help of a web tool, are compared to the results found in user testing. The case study thus serves as an example of how the framework can be used.

1.1. Tool vs. paper

In our previous empirical studies on heuristic evaluation (Law and Hvannberg, 2004a), evaluators complained that

Abbreviations: AE, actual efficiency; DV, dependent variable; HE, heuristic evaluation; ICT, Information and Communication Technology; IV, independent variable; PUP, predicted usability problems; SUPEX, Structured Usability Problem EXtraction; SUS, System Usability Scale; UAF, User Action Framework; UP, usability problem; UT, user test.

* Corresponding author. Tel.: +354 525 4702; fax: +354 525 4937.
E-mail addresses: ebba@hi.is (E.T. Hvannberg), law@tik.ee.ethz.ch (E.L.-C. Law), marta@ru.is (M.K. Lárusdóttir).

0953-5438/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2006.10.001

reporting problems on paper was cumbersome and time-consuming. This motivated us to attempt to improve the evaluation method by providing evaluators with a web tool. We wanted to improve the problem descriptions with further characterization of usability problems, such as context, cause, severity and relevant usability heuristics. The ultimate goal is to advance the validity of predictive methods, such as heuristic evaluation, with respect to user testing. In other words, to try to predict serious problems so that they are corrected in revisions of the user interface and to try to minimize the number of serious problems falsely predicted. We can reason that reporting usability problems using a software tool may help due to the following qualities:

- More accessible explanation of usability heuristics along with concrete examples.
- Easier to search, review, modify problem descriptions and link to relevant material (Gladwell, 2002).
- Faster entry of usability problems, thus making it more efficient.

In addition, the following may improve immediate management of usability problems:

- Merging of problem sets from different evaluators to get a unique set of problems.
- Measuring reliability of usability problems, i.e. whether one or more evaluators report them.
- Prioritizing usability problems according to impact, e.g. severity and cost of removal; tracking them through revisions, etc.
- Locating problematic contexts or tasks in the application, followed by designing task scenarios for user tests to evaluate these problems.
- Associating problems with previously proven patterns of solutions.

On the other hand, a software tool can have a negative effect on the usability problem reporting:

- Switching back and forth between the application being tested and the software tool can decrease the sense of context in the application for the evaluator. On the other hand, users frequently use a number of software applications in their work so such context-switching is common.
- More noise in the problem descriptions because of easy reporting, i.e. more False Alarms.
- Bias towards certain values in classification because of default values or order of values presented in menus.

So far, only a few software tools have been developed to support usability problem analysis, classification and reporting. A tool to assist evaluators with cognitive walkthrough and to record the results has been shown to improve the evaluation process over paper walkthrough (Rieman et al., 1991). The User Action Framework (UAF) (Andre et al., 2001) has a number of tools, including a usability design guide to be used during interaction design and usability lab testing, usability problem inspector to be used during formative evaluation, usability problem classifier and usability problem database. This database can provide valuable input to project management, solutions to problems, guidelines and relevant on-line literature. Emerging holistic frameworks, such as UAF, which can potentially support evaluators in using more than one method, can prepare them with appropriate training, help them with analysis, problem tracking and management. To the best of our knowledge, there has not been any empirical study, which compares the effectiveness of paper-based vs. tool-based usability problem reporting.

1.2. Nielsen vs. Gerhardt-Powals

Numerous sets of heuristics can be applied during heuristic evaluation (Folmer and Bosch, 2004). Many of them have common factors, such as consistency, task match, appropriate visual presentation, user control, memory-load reduction, error handling and guidance and support. Nielsen's heuristics have resulted from studies of practical applications in various contexts (Nielsen, 1993; Nielsen and Molich, 1990) but there is a lack of a sound theoretical framework to explain how they work. The less well-known principles, which were put forward by Gerhardt-Powals (1996), and are based on situation awareness and cognitive theory, have proven useful in a dynamic application such as anti-submarine warfare. Nielsen's heuristics are synthesized from a number of guidelines. Alternative sets of usability heuristics have been developed and tested. Most of the heuristics are design guidelines and refer to the system's user interface, and only a few heuristic sets are based on the understanding of user cognition (cf. Norman's theory of action model (Norman, 1986)) or situation awareness. Gerhardt-Powals put forward a set of guidelines based on cognitive principles. These guidelines are based on theory, but have not been adequately evaluated in practice. Specifically, we hypothesize that principles derived from cognitive engineering, which are strongly rooted in theories of cognitive psychology and other related disciplines, can well serve as a promising tool for heuristic evaluation.

Several factors, besides the set of usability heuristics, can influence the performance of heuristic evaluation, such as evaluator training, evaluator knowledge of the application domain, task coverage, problem extraction/description, merging, etc. A number of attempts have been carried out to improve heuristic evaluation, including the inspective, descriptive and analytical part. To improve the descriptive and the analytical part, Cockton and Woolrych (2001) have suggested a problem-reporting form and a way to analyse more accurately research procedures which can accurately count the number of problems discovered.

Evaluator's selection of tasks is an important part of a user test. Model-based evaluation has supported this in part, for example, by letting a task model guide the evaluator through the application. As most surveys (Rosenbaum et al., 2000) on usability methods indicate, practitioners apply more than one method, some of which are orthogonal to one another, such as paper prototyping and heuristic evaluation, and some of which complement one another like expert reviews and user testing. The discount usability engineering method (Nielsen, 1993, p. 17) uses four techniques: user and task observation, scenarios, simplified thinking aloud and heuristic evaluation. Nielsen (1994b, p. 58) suggests that there are two major reasons for alternating between heuristic evaluation and user testing. First, heuristic evaluation does not require users, who can be hard to get, and can be used initially to rinse out a number of usability problems. Second, a number of studies have shown that the two methods, heuristic evaluation and user testing, find distinct sets of problems. Frøkjær and Lárusdóttir (1999) conducted an empirical study, which showed that performing heuristic evaluation (HE) prior to user tests on the same system could somehow help non-expert evaluators uncover more usability problems in user testing, especially the severe ones. Such observation was derived from comparing the results of the two user tests, which were performed with and without doing HE beforehand. Besides, this combination of evaluation methods could eliminate one of the most important weaknesses of HE when used by non-experts, the proneness to addressing many false problems.

1.3. Task selection

One reason for a large number of predicted problems being False Alarms, i.e. not confirmed in user testing, is that users may not have been instructed to carry out the appropriate tasks and their set of tasks covered different contexts from the contexts that evaluators visited during heuristic evaluation. Thus, the issue of coverage of the application is of concern, when applying user tests to validate heuristic evaluation by comparing the respective lists of usability problems generated by these two UEMs (Usability Evaluation Methods). In heuristic evaluation, the portion of the application covered is up to the evaluator, especially if he/she is given total freedom of what aspects to cover. In a study of discovery methods, Cockton et al. (2003b) report that most evaluators choose to use system-searching or system-scanning more often than goal-playing and method-following. In a think-aloud user-testing, the coverage is influenced by the set of tasks presented to the user. Comparing two independent usability problem sets from two UEMs has drawbacks. An evaluator may predict problems that are miscoded as False Alarms since a user never had an opportunity to see them during user testing because he or she did not visit that part of the application (Cockton and Woolrych, 2001; Cockton et al., 2003b, 2004; Woolrych et al., 2004). Cockton and Woolrych (2001) suggest to systematically derive task sets for user testing from the initial set of predicted problems identified in heuristic evaluation so as to increase the power of user testing for exposing all predicted problems that really exist, i.e. eliminating instances of 'genuine' False Alarms. Another reason for the small overlap of predicted and actual usability problems is that the predicted problems state *causes* of problems but usability problems reported during user testing are frequently described as *effects* on the users (Cockton and Woolrych, 2001; Doubleday et al., 1997).

In view of the above discussion, we put forward in this paper the following research questions:

1.3.1. Tool vs. Paper

Do we achieve benefits in increased effectiveness, efficiency and inter-evaluator reliability in terms of a higher number of real usability problems in a shorter period by using a software tool to report predicted problems over using paper?

1.3.2. Nielsen vs. Gerhardt-Powals

Are Gerhardt-Powals' cognitive engineering principles more effective than Nielsen's usability heuristics in enabling evaluators to identify a higher number of real usability problems (higher validity) in a shorter time (higher efficiency)?

1.3.3. Task selection

Can the validity of heuristic evaluation be increased by using the set of usability problems so identified to guide the task selection of user tests whose results are in turn used to validate the outcomes of heuristic evaluation?

2. Materials and methods

Two experiments are described in this paper. A web portal called EducaNext (www.educanext.org) was evaluated in both experiments. In the first experiment, the portal was evaluated with heuristic evaluation with two sets of usability heuristics and two ways of reporting usability problems. In a two by two between-subject experimental design involving five evaluators in each of four cells, labelled A–D, we collected qualitative and quantitative data of predicted usability problems discovered during heuristic evaluation. The four groups are A, C (Nielsen heuristics), B, D (Gerhardt-Powals principles), where A and B used paper forms and C and D used the web tool for problem reporting.

Next, 8 task scenarios were designed based on the results from heuristic evaluation. In the second experiment, EducaNext was evaluated in user tests with 10 participants solving the task scenarios to check how many of the predicted usability problems (PUP) reported during heuristic evaluation were experienced by users as real problems (UP). We describe in detail how the predicted usability problems are filtered and matched with the usability problems discovered in the user test. The task selection process

for the user test is also described. The two experiments are described subsequently and an overview of the workflow is given in Fig. 1.

2.1. Heuristic evaluation

Members of higher education, research organizations and professional communities share, retrieve and reuse learning resources in a web portal, called EducaNext. EducaNext fosters collaboration among educators and researchers, allowing them to participate in knowledge communities; communicate with experts in their field; to exchange learning resources; to work together on the production of educational material e.g. textbooks, lecture notes, case studies and simulations; to deliver distributed educational activities: lectures, courses, workshops and case study discussions and to distribute electronic content under license. Nineteen Computer Science students in their final year of their BS-degree studies and one BS Computer Science graduate evaluated EducaNext using heuristic evaluation (HE). They had good knowledge of usability evaluation but little practice. Ten evaluators evaluated the portal using heuristics from Nielsen (1993), and 10 using heuristics from Gerhardt-Powals (1996) as a basis for the heuristic evaluation. Furthermore, usability problems were reported in two ways, 10 evaluators used a paper form and 10 used a web tool specially made for reporting problems for heuristic evaluation. Hence, there were five evaluators in each of the four groups, as seen in Table 1. In summary, two independent variables are Medium of Reporting (Paper vs. Tool) and Set of Heuristics (Nielsen vs. Gerhardt-Powals). A 2×2 between-subject factorial design was employed.

All evaluators were asked to evaluate EducaNext independently. They received email containing a checklist of the activities they were asked to perform. The instructions included a pre-evaluation questionnaire, a post-evaluation questionnaire, and an introduction material in a digital

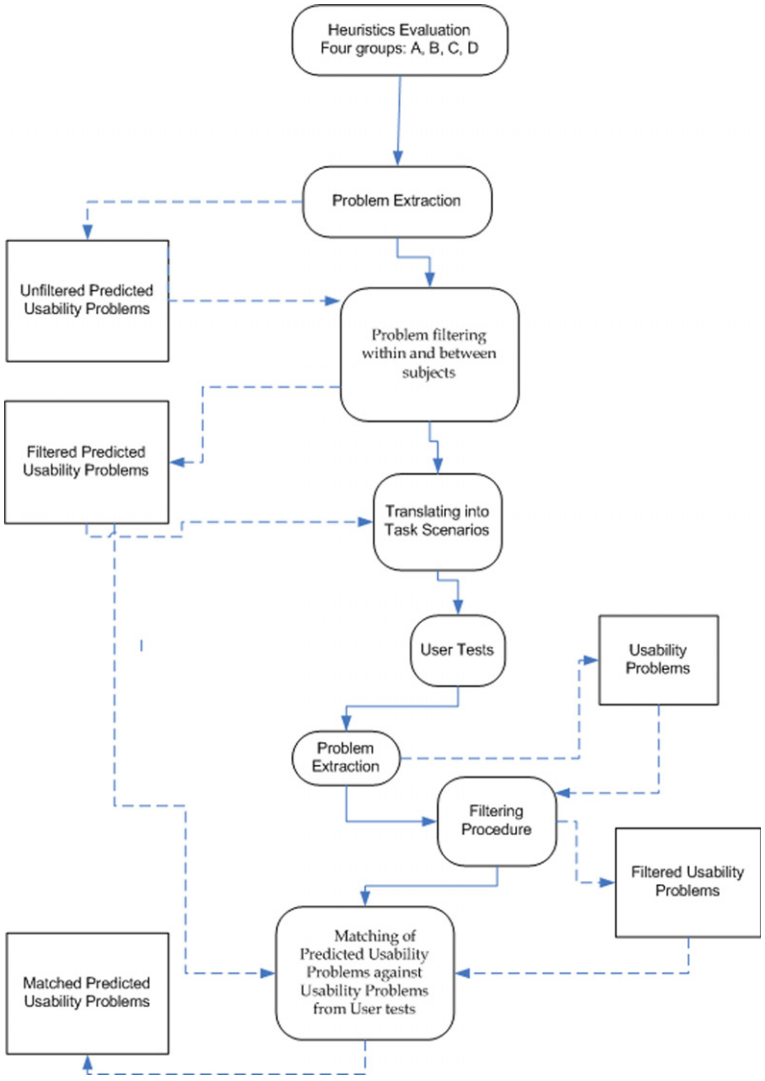

Fig. 1. Workflow of experiment.

Table 1
The experimental design of the Heuristic Evaluation

	Nielsen	Gerhardt-Powals
Paper	Group A: 5 novice evaluators	Group B: 5 novice evaluators
Tool	Group C: 5 novice evaluators	Group D: 5 novice evaluators

audio-file and slides including: (a) guidelines for the procedure of the evaluation; (b) introduction to the heuristic evaluation; (c) introduction to the EducaNext system and (d) introduction to the reporting of the usability problems. In the post-evaluation questionnaire, evaluators were asked to report the time spent, in hours, on the evaluation, and were asked to give a list of facilitators and hindrances of the heuristic evaluation method that they had applied.

2.2. Medium of reporting

Evaluators reported the usability problems in one of two ways, using a web tool (see Fig. 2) or on a paper form. A structured problem report format adapted from Cockton and Woolrych (2001) was used to report seven attributes of each predicted usability problem (PUP) (Table 2). Noteworthy is that the use of the structured problem report format can improve the reliability of merging PUPs and the reliability of matching predicted to actual problems (Cockton et al., 2003a), thereby increasing the overall internal validity of the usability evaluation results. The same attributes were reported on paper and with the tool, and the design of the forms on both media was similar. Heuristics used and severity rate were selected with combo boxes and a default value was given in both cases. Examples for reporting problems were provided on paper for the paper form, but in the web tool, evaluators could get help on attributes and a complete example usability problem. A list of all the heuristics and a list of the levels of severity, with further explanations via tooltips, were given in the tool, but the evaluators using the paper form were encouraged to print these out and have in front of them during evaluation.

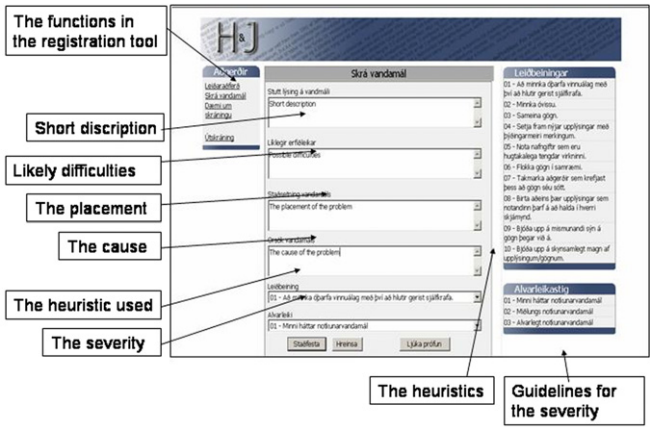

Fig. 2. User interface of a web tool for reporting usability problems.

Table 2
Structured problem report format (adapted from Cockton and Woolrych (2001))

1. A numeric identifier of the problem
2. A short description of the problem
3. Likely difficulties for the user
4. Specific context (the location of the problem in the interface)
5. Possible causes of the problem (what is wrong in the design)
6. The heuristic(s) used
7. The severity rate, containing 3 levels: severe, moderate and minor

The people using the web tool for reporting problems received a short introduction to the tool. They were recommended to use two computers, one for keeping EducaNext maximized and a laptop for reporting the usability problems in parallel.

2.3. Translating predicted problems into task scenarios

According to Cockton et al. (2003a), usability inspection methods (UIMs), such as heuristic evaluation, can serve as 'discovery resource' for user testing, which are designed to focus on potential problems predicted by UIMs, thereby improving construct validity (Gray and Salzman, 1998). Furthermore, Cockton and his colleagues (2004) demonstrate how UIM predicted problems could be translated into task scenarios of user testing. As the goal of falsification testing is to maximize confidence in false positive coding, definition of task sets for user tests should systematically be derived from UIM analysts' predictions. A task definition methodology essentially consists of three procedures: processing predictions, translating predictions into tasks, and verifying tasks against predictions. Here, we delineate how these three procedures have been implemented in our studies.

2.3.1. Processing PUPs

A usability specialist (E1), who is highly knowledgeable about the system tested, first examined closely each of the PUPs discovered by individual evaluators to discard within-evaluator duplicate or incomprehensible PUP descriptions. Then, E1 applied the 'problem reduction method' described in Connell and Hammond (1999) to filter out any overlapping PUPs to generate a list of unique PUPs. E1 then grouped these PUPs according to the attribute 'Specific Context' in the standard problem report form (Table 2), resulting in 15 groups, e.g. 'Access Content', 'Advanced Search', 'Simple Search', 'Browse Catalogue', 'Content Provision', 'Left Navigation Bar', etc.

2.3.2. Translating PUPs

With reference to her experiences about the usages of the system, E1 assessed the severity of individual PUPs and prioritized them within each group. PUPs with high priority were translated into task scenarios by abstracting the actions leading to their discovery (cf. the attribute 'Short Description' in the structured problem report).

Table 3
List of eight tasks for the user-test

Task 1: Apply for an EducaNext Portal User Account
Task 2: Login and Edit User Preferences
Task 3: Browse the Catalogue of the EducaNext Portal
Task 4: Simple Search
Task 5: Advanced Search
Task 6: Check the Booking History
Task 7: Create and Join an EducaNext Community
Task 8: Provide and Offer Educational Material

Caution was exercised to avoid over-constraining a user's action with too detailed task descriptions. As the 15 groups of PUPs are somewhat interrelated, one task scenario can address more than one group of PUPs. Finally, eight task scenarios (Table 3) with a set of sub-tasks addressing different PUPs were developed.

2.3.3. Verifying task scenarios

It was critical to check the coverage of the task scenarios to see whether any significant PUPs were left out from further validation. Another usability specialist (E2) mapped each of the unique PUPs against the eight task scenarios and found that 15 PUPs were not covered by any of the scenarios. E1 repeated the same mapping exercise and identified two mapped cases that E2 interpreted differently. E1 and E2 then negotiated to assess the relevance of the omitted PUPs. Consequently, they decided to ignore eight of these PUPs, which were highly situational (e.g. depending on the type of the content that the user retrieved from the portal) or trivial, and incorporated the other five into the existing task scenarios.

2.4. User tests

The user tests (UT) were conducted to find out how many of the predicted problems reported by the evaluators using the heuristic evaluation were real problems for users. The task sets for the user tests were derived from the results of the heuristic evaluation, asking the users to use those features of EducaNext that were found to be problematic in the heuristic evaluation (Section 2.3). This method was adopted to increase the power of the user tests for validating the predicted usability problems.

Ten participants ($P_1, \ldots, P_{10}$) were asked to solve eight tasks (Table 3) while thinking aloud. First, the participants answered a pre-test questionnaire about their personal and technical background. Then, they solved the task scenarios one by one and answered a so-called 'After Scenario Questionnaire' (Lewis, 1991) in between measuring their subjective satisfaction about that task. Everything that happened on the screen was captured using a screen capture tool and a web camera was used to record both sound and video of the session. After completing the eight tasks, the users were asked to answer a post-test questionnaire SUS (System Usability Scale) (Brooke, 1996) measuring their subjective satisfaction about EducaNext as a whole. SUS has been proved a robust, reliable, and low-cost usability assessment tool that is widely used in industry. The 10 participants were all university staff, either professors or administrators. There were five males and five females. Two participants were in their thirties, seven participants were in their forties and one in her fifties. The average ICT (Information and Communication Technology) competence was self-assessed at 3.5 (out of 5, $N = 10$) ($SD = 0.97$). Participants reported an average of 3.5 (out of 5, $N = 10$) grade when asked about experience in e-learning ($SD = 1.27$). Seven of the participants had developed or organized online learning content.

An experimenter observed the users unobtrusively, in the same room, while they were performing the task scenarios. For each task, she reported the user's performance (start-time, end-time, number and type of errors, instance and type of help seeking, instance of expressed frustration), comments and problems with a template on a laptop. The experimenter checked these data against the screen captures and audio recordings to ensure their accuracy.

2.5. Problem extraction, filtering and matching procedures

The inability to expose all actual problems is recognized as an inherent limitation of user testing. Notwithstanding, user test results are employed to validate UIMs. Errors in generating a set of actual problems, identified in user tests, may lead to miscoding predicted problems as False Alarms. The SUPEX (Structured Usability Problem EXtraction) method (Cockton and Lavery, 1999) addresses this issue of problem extraction. However, the *usability* of this method, especially the learnability, is questionable because of its involved stages and sub-stages. The efficiency of applying SUPEX appears low, given the time-consuming processes such as segmentation, transcription and coding. The cost-effectiveness of the method is not clear. Depending on the budgetary constraint and other contextual factors, stages and sub-stages of SUPEX can be skipped. What is left, when stripping the method to its core elements, will be a common approach to problem extraction, which is more or less the same as the one adopted in the current study. Such flexibility in applying the method makes it difficult to decide between SUPEX and non-SUPEX, and to prove its claimed advantages. The comprehensive assessment and development of SUPEX is yet to be done, especially its usability and reliability. Consequently, we relied on the traditional approach to problem extraction from the user tests.

The usability specialist E1, who has performed several user tests on different versions of EducaNext, became very knowledgeable about the system, which was vital for effective and efficient problem extraction. Specifically, she referred to the conventional definition of usability problem to guide the extraction task, namely:

> *"Simply stated, a usability problem is any aspect of a user interface that is expected [or observed] to cause users*

problems with respect to some salient usability measure (e.g. learnability, performance, error rate, subjective satisfaction) and that can be attributed to a single design aspect" (Nielsen, 1993, p. 388)*, [our addition].*

For individual users, E1 derived a list of usability problems from their think-aloud protocols and from the experimenter's detailed observation notes. Furthermore, the 10 lists of usability problems were merged and the overlapping ones filtered out. Problem instances rather than problem types were counted (John and Mashyna, 1997). In other words, the same problem identified in two different contexts would be counted as occurring twice rather than once. Observed UPs were recorded in the same structured report format used for PUPs, sharing the same set of attributes, including Identifier, Users/Evaluators Involved, Description, Context, Severity and Frequency, thereby facilitating the matching task.

In addition, E1 devised a so-called two-way mapping procedure (forward- and backward-matching) for validating the PUPs of heuristic evaluation against the UPs of user tests. First, we took PUP1 and mapped it to each of the 58 UPs; then repeated the same procedure for PUP2 up to PUP85. Then, we took the UP1 and mapped it to each of the 85 PUPs; then repeated the same procedure for the remaining UPs. With this approach, the reliability of the mapping results could be enhanced, though it was tedious and time-consuming. Fig. 1 illustrates the aforementioned procedures.

3. Results

In the following three subsections, we report on the results of the heuristic evaluation experiment, user tests and the validity and thoroughness of heuristic evaluation compared to the user tests after the usability problems of the two experiments have been filtered and compared. We, thus, report how well the evaluators conducting heuristic evaluation were able to predict usability problems reported by the users. Finally, we compare two and two groups together, i.e. on the one hand, those that apply Nielsen vs. Gerhardt-Powals and those applying paper vs. tool, on the other hand. To compare the different groups, we calculated the thoroughness, validity and efficiency of the evaluation. The any-two-agreement measures of the evaluators (Hertzum and Jacobsen, 2001) within the groups were evaluated.

3.1. Data analysis from heuristic evaluation

During heuristic evaluation, there were altogether 160 predicted usability problems (PUPs) reported by 20 evaluators. Note that no PUP was reported by B1 (see Table 4). Groups A, B, C and D reported 33, 37, 43 and 47 PUPs, respectively. χ^2 test showed that there were no significant differences in the number of PUPs between the four groups. Examining means of the four groups revealed that there was close to significant difference at the 0.05 level in the number of PUPs between Group A and Group C ($t = 2.21$, $df = 8$, $p = 0.06$), i.e. between paper and tool registration when using the Nielsen's heuristics, but no difference when using the Gerhardt-Powals' principles.

Table 4
Number of predicted usability problems (PUPs) per evaluator in each group

		Nielsen Heuristics					
Paper	ID	A1	A2	A3	A4	A5	$M = 6.6$
	PUPs	6	5	6	9	7	$SD = 1.4$
Tool	ID	C1	C2	C3	C4	C5	$M = 8.6$
	PUPs	10	10	8	8	7	$SD = 1.2$
		Gerhardt-Powals Principles					
Paper	ID	B1	B2	B3	B4	B5	$M = 7.4$
	PUPs	0	10	8	8	11	$SD = 3.9$
Tool	ID	D1	D2	D3	D4	D5	$M = 9.4$
	PUPs	4	8	11	13	11	$SD = 3.1$

Note that B1, who apparently was not motivated for the given task, uncovered zero PUPs. When that subject was removed from the population, analysis further showed that there was significant difference in the number of PUPs between Group A and Group B ($t = -2.62$, $df = 7$, $p = 0.04$), i.e. between Nielsen and Gerhardt-Powals when using paper for problem reporting, with the latter discovering more usability problems than the former.

A usability specialist examined each of the PUPs and discarded 12 of them, which were reported more than once by the same participant or were found to be incomprehensible. Consequently, 148 out of the original 160 PUPs were consolidated to eliminate any duplicate (Section 3.2), using the procedure described in Connell and Hammond (1999). This exercise filtered out duplicates and led to a list of 85 unique instances of PUPs, on which the design of task scenarios of a user test (UT) were based; these PUPs were validated by the results of the user test thus developed (see below).

Hertzum and Jacobsen (2001) developed a metric known as "Any-Two Agreement" (see Eq. 1) to estimate inter-evaluator reliability, i.e. the probability that a particular usability problem is identified by more than one evaluator. This measure is supposed to be more accurate than the conventional method of using simple problem discovery rate (p), which can be much inflated when a sample size is small.

$$\text{Any-two-agreement} = \text{Average of } |P_i \cap P_j| / |P_i \cup P_j| \quad (1)$$

over all $1/2\ n\ (n-1)$ pairs of users, where P_i and P_j are the sets of UPs identified by user$_i$ and user$_j$, and n is the number of users.

We computed Any-two-agreement for all the 20 evaluators of heuristic evaluation as a whole, for each of the four groups and for a combination of the groups sharing a common feature. The overall Any-two-agreement for the 20 evaluators was 0.043 (or 0.046 for the 19 evaluators excluding B1). The results are shown in Table 5, with the numbers in parentheses including evaluator B1.

Table 5
Inter-evaluator reliability measured with any-two-agreement

	Nielsen	Gerhardt-Powals	
Paper	A = 0.042	B = 0.088 (0.053)[a]	A + B = 0.045 (0.041)
Tool	C = 0.073	D = 0.070	C + D = 0.055
	A + C = 0.047	B + D = 0.046 (0.034)	All = 0.046 (0.043)

[a] Numbers in parenthesis include evaluator B1.

The overall inter-evaluator reliability was very low, or 0.046, in comparison with the average Any-two-agreement value of 0.101 within the group of user-test participants. The low reliability can be attributed to the heterogeneous backgrounds of the evaluators, especially their experience and knowledge of usability evaluation and of the domain of the system evaluated (i.e. e-learning). Besides, the two independent variables (i.e. set of heuristics and type of medium for problem reporting) had certain effects on the nature and number of usability problems identified. Furthermore, the differences between the groups were not statistically significant. Group C showed the highest level of agreement. It may be explained by the two factors: the evaluators' familiarity with Nielsen's heuristics and the facilitating effect of the tool. Nonetheless, due to the small number of participants, these assumptions cannot be confirmed.

After coding the qualitative data on facilitators and hindrances the evaluators mentioned, we obtained the results listed in Table 6. It is interesting to note that the heuristic set was considered as both a facilitator ($n = 6$) and a hindrance ($n = 7$), and that Gerhardt-Powals' were perceived as more hindering than were Nielsen's. This finding raises the issue whether evaluators, without being given any guidelines (i.e. control), could outperform those with any of the two sets. At least two studies, comparing guidelines to a control group, where the control group does not rely on any guidelines but only their own knowledge, have been performed in the past (Bastien et al., 1999; Connell and Hammond, 1999). In the study by Bastien et al. (1999), where median number of usability problems uncovered by the participants was used as a metric, the control and the ergonomic criteria groups differed significantly, with the latter uncovering more problems but also spending more time on the evaluation. In the same study, significant differences did not appear between the control group and a group using ISO/DIS 9241-10 dialogue principles. In the study by Connell and Hammond (1999), where Nielsen's heuristics and a set of 30 principles were compared to a control group, no significant differences were found between the conditions for a group of novices. That more evaluators perceived training as a facilitator than a hindrance implies that the training given was somewhat effective and desirable. Three of the 20 evaluators thought that they did not have enough time to evaluate the application. This may indicate that the task was more time-consuming than they expected and could have been broken down to more than one assignment. It raises the question whether the number of usability problems discovered is highly dependent on the amount of time spent. The Pearson correlation showed that there is a moderate relationship between the two variables ($r = 0.48$, $p = 0.03$).

Table 6
Hindrances and facilitators

Hindrance	Freq.	Facilitator	Freq.
Unclear training/still not a problem	1	Training Material	8
Heuristic guidelines	7	Heuristic guidelines	6
Lack of experience	1	Previous experience	1
Lack of time	3	Using two computers	1
Difficult application/ EducaNext	3	Easy application/ EducaNext	1

3.2. Analysis of usability problems from the user tests

There were altogether 125 unconsolidated usability problems (UPs) experienced by the 10 participants in user testing. The average number of problems reported per participant during user testing was $M = 12.5$, $SD = 4.2$, $N = 10$, which was higher than that of heuristic evaluation ($M = 8.0$, $SD = 2.9$, $N = 20$).

Because of the filtering procedure, a list of 58 UPs was produced, of which 10 were severe, 27 were moderate and 21 were minor. Besides, only one participant experienced 24 of these 58 UPs and six participants experienced only one UP, which was severe (see Fig. 3). This list was used to validate the PUPs reported in heuristic evaluation.

The average problem discovery rate (p) over the 10 participants was 0.22, which was not particularly high. In other words, at least seven users were required to identify 80% of the discoverable UPs (cf. the assumption of "Magic Number 5" (Barnum, 2003) – *five* users can yield 80% of the findings from a usability test – also supported by Nielsen (2000) and other human factors engineer (e.g. Virzi, 1992) – but questioned by several researchers (Faulkner, 2003; Law and Hvannberg, 2004b; Spool and Schroeder, 2001; Woolrych and Cockton, 2001)). Furthermore, the problem discovery rate (p) for severe problems was 0.3, which was higher than that of moderate ($p = 0.21$) and that of minor problems ($p = 0.18$).

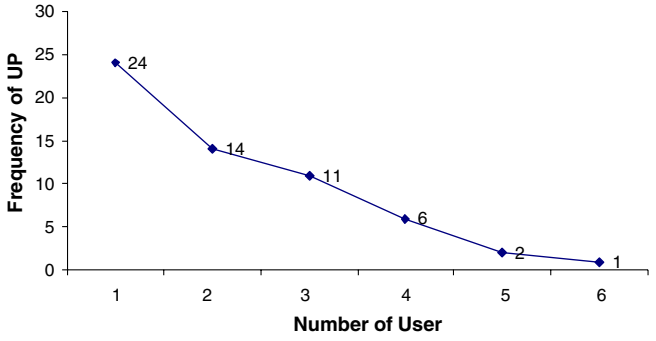

Fig. 3. Number of users experiencing UPs with different frequencies.

The 10 participants were required to fill out an After-Scenario Questionnaire (ASQ) for each of the eight scenarios to assess the perceived ease, perceived time and perceived utility of online help. Each of these three aspects was measured with the seven-point Likert Scale with the left and right anchor being "Strongly Disagreed" and "Strongly Agreed", respectively. Furthermore, upon attempting all the tasks, the participants were required to complete a post-test questionnaire (SUS) to assess their overall subjective satisfaction with EducaNext. The average SUS score was 57.25 ($N = 10$, $SD = 16.43$). Note that SUS scores have a range of 0 to 100 (Brooke, 1996). For each of the 10 participants, we computed two different objective measures: *UP-severity-ratings* and *total-time-on-tasks*. Specifically, we added up the severity ratings (severe = 3, moderate = 2 and minor = 1) of all the UPs that a participant experienced, resulting in his or her UP-severity-ratings. Similarly, we added up the time that a participant spent on all the tasks attempted. The average UP-severity-rating was 23.6 ($N = 10$, $SD = 7.79$) and the average total-time-on-tasks was 54.6 min ($N = 10$, $SD = 12.04$).

3.3. Mapping the results of heuristic evaluations with those of user tests

Using the two-way mapping procedure described in Section 2.5, the following results were obtained:

Number of Hits (i.e. PUPs verified by UPs) = 32

Number of False Alarms

(i.e. PUPs not verified by any UP) = 53

We studied how many of Hits and False Alarms were identified by different groups of HE evaluators. Results are displayed in Table 7. Note that within-group duplicates were not counted, yielding so-called 'filtered Hits'. For instance, PUP20 was a hit uncovered by A1, A3, C1 and C5; however, the corresponding frequency of hit was only increased by 1 for Group A and for Group C. χ^2 tests were applied to these results. On average, an evaluator in the heuristic evaluation demonstrated similar performance in terms of Hits and False Alarms, irrespective of the types of supports given, i.e. heuristic sets and reporting medium. Compared to the results of heuristic evaluation, the group that used Nielsen's heuristics and reported on paper (Group A) identified on average 6.6 usability problems per evaluator, but on average only 2.6 of these problems per evaluator could be verified in the user test.

Furthermore, to find out whether tool-based reporting was more effective than reporting on paper, we collapsed the data of the respective groups (Table 7) and performed *t*-tests between Group A + Group B and Group C + Group D. No significant differences in the number of Hits or False Alarms between tool-based and paper-based evaluator were found. Similarly, to find out whether Nielsen's usability heuristics were more effective than Gerhardt-Powals principles, we collapsed the data of the respective groups and performed *t*-tests between Group A + Group C and Group B + Group D. No significant differences were found either. In summary, on average, an evaluator in the heuristic evaluation demonstrated similar performance, irrespective of the types of supports given.

Besides, it is intriguing to know how many *severe* verified UPs and *unique* verified UPs (i.e. UPs were reported only by the evaluators belonging to one particular group) individual groups identified. Table 8 displays the results. Group B appeared to be the most effective in identifying severe UPs and Group B and C in identifying additional UPs, but the differences were insignificant.

In addition to Hits and False Alarms, misses are denoted as those usability problems that inspection methods have missed but are found in user testing. We computed the overall effectiveness of heuristic evaluation based on the following formulae (Hartson et al., 2001): Effectiveness = Validity * Thoroughness, where

$$\text{Validity} = \text{Hits}/(\text{Hits} + \text{False Alarms}) = 32/85 = 0.38 \quad (2)$$

$$\text{Thoroughness} = \text{Hits}/(\text{Hits} + \text{Misses}) = 32/58 = 0.55 \quad (3)$$

$$\text{Overall Effectiveness} = 0.38 * 0.55 = 0.21 \quad (4)$$

In addition, we researched whether the HE evaluators and UT users tended to identify the real UPs with a similar frequency. To answer this question, we correlated the two frequencies of individual Hits. The non-parametric Spearman Rank-Order Correlation Coefficient r_s was 0.23 ($p = 0.2$, $N = 32$), showing that the two groups, HE evaluators and UT users, were not significantly correlated in this respect.

Table 7
Distribution of *Hits* and *False Alarms* – filtered sets

	Hits		False Alarms	
	Nielsen Heuristics	Gerhardt-Powals principle	Nielsen Heuristics	Gerhardt-Powals principle
Paper	Group A = 8	Group B = 13	Group A = 18	Group B = 12
Tool	Group C = 15	Group D = 11	Group C = 19	Group D = 26

Table 8
Distribution of *severe* verified problems and *unique* verified problems

	Severe problems		Unique problems	
	Nielsen Heuristics	Gerhardt-Powals principle	Nielsen Heuristics	Gerhardt-Powals principle
Paper	Group A = 2	Group B = 5	Group A = 2	Group B = 6
Tool	Group C = 2	Group D = 2	Group C = 8	Group D = 4

3.4. Results according to heuristics

To answer the question which of the two sets of heuristics can result in higher validity and thoroughness, we collapsed the data of the respective groups and analysed them.

We adopted the following definition (Law and Hvannberg, 2004a):

$$\text{Actual efficiency(AE)} = \text{Number of Hits identified within the Testing Session/Total duration(hours) of the Testing Session} \quad (5)$$

From Table 9, we see that the validity is almost the same for the two sets of heuristics, i.e. less than half of the predicted problems could be verified in the user tests. (Note that if a problem was discovered with Nielsen's heuristics either with paper or the tool, it was only counted once.) If the results of the user tests are the baseline, i.e. the truth that we can compare to, the measures indicate that about 60% of the effort of doing heuristic evaluation has been wasted. Similarly, the thoroughness for the two sets was almost identical. The measures indicate that more than 60% of the usability problems discovered by the users were undetected by the evaluators in the heuristic evaluation.

The usefulness of a usability evaluation method depends in part on whether it enables the evaluator to discover the most severe problems and whether it avoids misleading developers to fix False Alarms (i.e. those reported problems not verified by the user tests). The distribution of problem severity between heuristic sets in Table 10 was not significant.

The average time it took each evaluator to do the evaluation with Nielsen's set of heuristics (2 h 18 min) was only 5 min less than the average time it took an evaluator with Gerhardt-Powals principles (2 h 23 min).

The evaluators' satisfaction was mixed. Independent of heuristic sets, some said it was time-consuming to use the heuristics, there were too many heuristics and hard to understand, while others said they had no problems. Many mentioned that while evaluating, they first found a problem and then had a hard time finding the right heuristic to refer to. This may indicate that the heuristics are not always explicitly guiding the evaluators to discover problems or that the evaluators are finding problems for which there exist no heuristics in the respective set of heuristics. Relating to the first issue, i.e. guidance of heuristics in discovering a problem, in a study where an extended problem report format was used, Cockton et al. (2003b) found significant improvements in appropriateness scores increasing to a mean score of 61% from an earlier one of 31% (Cockton and Woolrych, 2001). They contributed at least part of the difference to discovery methods and part to the extended problem report format. Cockton and Woolrych (2001) have termed inappropriate heuristics when an expert rather than a heuristic evaluation is being applied. Regarding the second issue, i.e. no heuristics existing, apparently, it is difficult to design a heuristic set, which has total coverage. Already, Nielsen (1994a) showed that in a factor analysis, only 30% of the variance was due to seven main heuristics factors which motivated him to relax the criterion of coverage from a problem matching a single heuristic perfectly (rating of 5) to a partial match (rating of 3). With this criterion, he assumed that usability problems are due to a broad variety of underlying phenomena.

Law and Hvannberg (2004a) show that not only is it sometimes difficult for the evaluators to find the matching heuristic, but they find it difficult to associate a problem with any of the heuristics. Their study concluded that evaluators *"identified quite a number of the severe UPs based on their own personal experiences and intuitions"* (Law and Hvannberg, 2004a, p. 247), and more so for the Gerhardt-Powals principles than for the Nielsen's heuristics.

3.5. Effectiveness and efficiency according to the medium of reporting

The medium of reporting could have some impact on how many usability problems are found during heuristic evaluation and how serious the problems are rated. We computed the validity, thoroughness and efficiency of the results for the two different media used to report the usability problems, paper or the tool, with 10 evaluators in each group.

Table 11 shows that the validity was almost the same for using a paper and using a tool, i.e. even though many more problems were reported using the tool than the paper. So almost 60% of the effort has been wasted using paper and about 55% of the effort has been wasted using a tool.

Table 9
Validity and thoroughness of heuristics problem sets

	Nielsen (A + C)	Gerhardt-Powals (B + D)
Validity	21/53 = 0.40	21/54 = 0.39
Thoroughness	21/58 = 0.36	21/58 = 0.36
Efficiency (AE)	21/21.3 = 0.98	21/23.3 = 0.90

Table 10
Verified problems according to problem severity and heuristic sets

	Nielsen (A + C)	Gerhardt-Powals (B + D)	User test
Severe	3 (30%)[a]	5 (50%)	10
Moderate	13 (48%)	10 (37%)	27
Minor	5 (24%)	6 (26%)	21

[a] (Validity) = % of verified problems against the total number of usability problems of respective severity identified in the user test.

Table 11
Validity of heuristic evaluation according to the medium of problem reporting

	Paper (A + B)	Tool (C + D)
Validity	18/44 = 0.41	23/51 = 0.45
Thoroughness	18/58 = 0.31	23/58 = 0.40
Efficiency (AE)	18/24 = 0.75	23/20.7 = 1.1

Table 12
Verified problems according to the problem severity and medium of reporting

	Paper (A + B)	Tool (C + D)	User Test
Severe	6 (60%)[a]	4 (40%)	10
Moderate	7 (26%)	15 (56%)	27
Minor	5 (29%)	4 (19%)	21

[a] (Validity) = % of verified problems against the total number of usability problems of respective severity identified in the user test.

The thoroughness for using a tool was slightly higher. The measure indicates that almost 70% of the usability problems experienced by the users were not reported by the evaluators in the heuristic evaluation using a paper form, and 60% of the usability problems experienced by the users were not reported by the evaluators in the heuristic evaluation using a tool to report the problems.

There was no significant difference in the distribution of severe, moderate and minor problems (see Table 12).

The average length of an evaluation session per evaluator using a tool was 2 h 14 min, which was 16 min less than using a paper (2 h 30 min).

The evaluators using the paper form stated that it would be tiring in the long run to use the paper form. The tool users were very positive, stating, that the tool was easy to use; it was effective to have tooltips and examples for guidance for reporting the problems; it was efficient to use drop-down lists to choose the severity rate and heuristics used to find the problem. They would rather use the tool than similar paper form.

4. Comparing two empirical studies

Previously, the authors performed an empirical study on estimating and improving the effectiveness of heuristic evaluation. The former study (Study 1) (Law and Hvannberg, 2004a) shared some similarities with the current one (Study 2), including:

- Employing Nielsen's heuristics and Gerhardt-Powals principles and applying them to an e-learning platform (NB: versions v.0.85 and v.1.0 of the same platform were tested in Study 1 and Study 2, respectively).
- Validating the results of heuristic evaluations with those of user tests.
- Comparable number of participants ($n = 4$ or 5) for each of the four groups.

In contrast, there are four major important differences (see Table 13) between the two studies, including:

1. More structured and detailed information about the system evaluated was provided to the evaluators in Study 2 than in Study 1.
2. Between-subject design was used in Study 2 instead of within-subject design in Study 1.
3. Using "Paper-based reporting vs. tool-based reporting" as one of the two independent variables in Study 2 instead of using "Textual descriptions versus Graphical representations" about the system tested in Study 1.
4. User tests were conducted *after* heuristic evaluations and were designed based on the data of heuristic evaluations in Study 2, whereas user tests were conducted *before* heuristic evaluations in Study 1.

The major finding in Study 1 was that Nielsen's heuristics could enable the evaluators to identify significantly more actual usability problems (Hits) than Gerhardt-Powals principles, irrespective of the type of the other supporting information the evaluator received. However, Study 2 could not corroborate this finding. It is intriguing to try to understand what may contribute to this contradictory conclusion. We examine the four factors delineated above:

Factor 1: the extra information given to the evaluators in Study 2 focused on the system evaluated but not on the heuristics or principles. Hence, assuming that the positive effect of increasing the understanding of the system was

Table 13
Comparison of the two studies

		Former (Study 1)	Current (Study 2)
Research design	Independent variables (IV)	IV1: Textual vs. Graphical description of the system	IV1: Paper- vs. Tool-based problem reporting
		IV2: Reference set – Nielsen vs. Gerhardt-Powals	
	Dependent variable (DV)	Number and quality of usability problems	
	Experimental design	Within-subject design (4 cells, each with 5 subjects)	Between-subject design (4 cells, each with 5 subjects)
	Pre-test training	Focus on the usability guidelines	Focus on the main features of the system
	Validation	Results of user tests [validate]	Heuristic Evaluation [guide] design of user tests
		Results of Heuristic Evaluation	Results of user tests [validate] Results of Heuristic Evaluation
Main findings	Form of support (IV1)	No effect on dependent variable	
	Heuristic set (IV2)	Significant difference: Nielsen more effective	No significant differences

more or less equal for all the four groups, it is rather unlikely that this factor can explain the insignificant difference between Nielsen and Gerhardt-Powals in Study 2.

Factor 2: We argued elsewhere for the legitimacy of using within-subject design in the context of Study 1 and also analyzed the inherent characteristics of Nielsen's heuristics that may render them more effective than Gerhardt-Powals' principles (Law and Hvannberg, 2004a).

Factor 3: Given the insignificant effects of the independent variable "textual vs. graphical" of Study 1 and "paper versus tool" of Study 2 in yielding more Hits or even more False Alarms, this factor should play no role in explaining the contradictory conclusion.

Factor 4: The special arrangements of the user tests for validating heuristic evaluations in Study 2 were to enhance the validity of the mapping results, and this purported effect should be more or less equal for all the four groups.

However, we cannot definitely eliminate the possibility that the significant differences found in Study 1 could be an artifact of the experimental design. Hence, Factor 4 may partially, but not fully, explain the insignificant differences found in Study 2. Apparently, the so-called user effect (Law and Hvannberg, 2004b) and evaluator effect (Hertzum and Jacobsen, 2001; Molich et al., 2004) can account for the contradictory finding between the two studies. In fact, evaluator effect can boil down to individual differences (Dillon and Watson, 1996) that are difficult to control in empirical studies in HCI.

Nonetheless, given the small number of evaluators per group, the effect size or power of the statistical tests employed in both studies is not particularly high. Clearly, an empirical study of a larger scale is required. Noteworthy is that the overall effectiveness of heuristic evaluation was disappointingly low in both studies – 0.22 for Study 1 and 0.21 for Study 2. This finding may threaten the claim that heuristic evaluation is an effective discount method for evaluating usability.

5. Discussion

Before we summarise the answers to the research questions we raised in Section 1 and suggest research issues for further work, we will point out the context of the study that will help us draw conclusions from the results.

Novice evaluators performed the heuristic evaluation in the first of the two experiments presented in this paper. As stated by Molich and Jeffries (2003), the former being one of the inventors of heuristic evaluation, it can be applied by *"someone without particular knowledge of usability engineering to evaluate a user interface"* (Molich and Jeffries, 2003, p. 1060). Hertzum and Jacobsen (2001, p. 424) also stated that: *"Any computer professional should be able to apply heuristic evaluation, but the informality of the method leaves much to the evaluator."*

Indeed, some of the gaps between the results of the heuristic evaluation and those of the user tests in this paper could be attributed to the evaluator effect (Hertzum and Jacobsen, 2001), i.e. levels of expertise and experience of individual evaluators have observable influences on usability evaluation outcomes. We tried to mitigate the evaluator effect by providing highly structured training material on heuristic evaluation and on the system. The evaluators got a checklist of the activities they were asked to perform and a digital audio-file introducing heuristic evaluation, the EducaNext system and the process of reporting problems. An analysis of the qualitative data on facilitators and hindrances of the heuristic evaluation showed that 8 of 20 evaluators found that the training material helped them, but one evaluator mentioned that training was a hindrance and one mentioned lack of experience.

In the following, we summarize the answers to the research questions we raised in Section 1 and put forward further research issues.

5.1. Tool-based vs. Paper-based Reporting

Part of this study is an attempt to answer whether tool-based reporting is superior to paper-based in terms of yielding a higher number of real usability problems (increased validity) within a shorter period of time (increased efficiency). The proposition is that typing can be faster than handwriting, digital content is more accessible and easily modifiable than its paper version and instructions are more readily available. We achieved small benefits of tool over paper, the effectiveness was 0.18 when using a tool compared to 0.13 using paper, but the difference was not significant. No significant differences in Actual Efficiencies per evaluator were found between Group A (AE: $M = 1.34$, $SD = 1.1$) and Group C (AE: $M = 2.1$, $SD = 2.2$) or between Group B (AE: $M = 1.32$, $SD = 0.9$) and Group D (AE: $M = 2.2$, $SD = 1.1$) or between Group A + B and Group C + D. While the evaluators using Nielsen's heuristics tended to type more with the web tool than to write with the conventional paper-and-pen ($t = 2.21$, $df = 8$, $p = 0.06$), it did not imply a higher quality of these PUPs. When asking participants to list hindrances and facilitators of heuristic evaluation, none of them mentioned the registration tool but one in group A mentioned the paper form as a hindrance. The ineffectiveness of our tool in enhancing the validity or efficiency can be attributed by the three negative effects addressed earlier in Section 1, including the cognitive load caused by switching between the two software systems, hasty data entry resulting in false alarms, and biased use of certain classification values.

We may need to implement more intelligence in the tool to gain further advantages. As it is, the tool provides a good basis for problem recording, providing help to its users that they reportedly liked. Although an earlier study (Law and Hvannberg, 2004a) showed no difference between those that used training material (Lavery et al., 1996) for heuristic evaluation and those who did not, the reason may have been that the training material was not at hand during evaluation but was exposed to the evaluator

before the analysis. Additional assistance to evaluators may be given in a tool by:

- Linking the usability problem better to the context in terms of screen scenarios or individual design features, for richer problem description and hopefully more efficient problem fixes.
- Helping evaluators go through falsification testing to mitigate, if not totally eradicate, instances of false alarms.
- Pointing out gaps in testing coverage of the application to avoid missed problems.
- Pointing to previously proven problematic areas that need to be retested (i.e. inspecting a revised version of the system).
- Giving them 'usability problems profile' that consists of problematic areas commonly identified for the type of products or interfaces under evaluation (Chattratichart and Brodie, 2004).
- Fostering the reliability and validity of problem severity ratings with automatable problem-rating rules derived from a robust theoretical and computational model (Blackmon et al., 2005) that are built in an intelligent reporting tool.

Research of the features and affordances of tool vs. paper has taken place in different domains, such as air traffic control, document handling (Gladwell, 2002) and user interface design (Cook and Bailey, 2005). Whereas paper has a unique set of affordances such as being tangible, spatially flexible and tailorable, digital documents can easily be searched, shared, stored, accessed remotely and linked to other related material (Gladwell, 2002). These affordances of paper, it is claimed, make paper attractive for creative collaboration and help maintain workers' mental models. In addition to these features, designers have reported that they find paper quicker, easier and more portable, in early designs (Cook and Bailey, 2005). Note that the difference between the representation of designers and evaluators of this study is that the latter produced text as deliverables but the former worked with sketches and text. We did not study the collaborative task of merging usability problems, but this may be an interesting subject for further study.

5.2. Nielsen vs. Gerhardt-Powals

Another approach to improving heuristic evaluation is to search for better heuristic sets. The validity of the two heuristic sets, Nielsen's and Gerhardt-Powals', with respect to user testing was the same. The overall effectiveness was 0.14 in both cases. Percentage wise, Gerhardt-Powals' principles could enable more severe problems to be found although the difference was not significant. Only looking at predicted problems, both the average number of problems and the standard deviation for Gerhardt-Powals' principles were higher than those for Nielsen's heuristics. This indicates that the Gerhardt-Powals' principles could further be exploited but that evaluators need more training. The findings of this study are contradictory to those of a previous study (Law and Hvannberg, 2004a), which implies that an empirical study of larger scale is needed.

The low effectiveness may be due to mismatch in the application domain knowledge between the evaluators in the two groups. Furthermore, in comparison with Nielsen's heuristics, the hypothesized strengths of cognitive engineering heuristics may not be exploited in a web-application with standard operations such as search, insert, access, and delete. The implication of this observation for future research is to compile and systematically evaluate a list of usability heuristics for e-learning systems. Indeed, there have been some recent studies in developing tailor-made usability heuristics to fit special application contexts (e.g. ambient displays, (Mankoff et al., 2003)) and large screen display information exhibits (Somervell et al., 2003), and they are proven to be more effective than Nielsen's heuristics (Somervell and McCrickard, 2005). A tool, with the added intelligence of creating taxonomies of problems in various application contexts to context specific heuristics, can make such a selection of heuristics more effective.

5.3. Task selection

The difficulty in creating tasks for user testing was twofold: the heuristic evaluation revealed that some of the problems were not reproducible at all from the problem descriptions, and some problems were situational and contextual, i.e. depending on particular data retrieved or entered by the evaluator, and on the stability of the server where the system evaluated resided.

A simple count of number of problem contexts, that can be defined as an identifiable place within the application, in each of the two problem sets – heuristic evaluation and user tests – shows that of the 23 contexts with usability problems in the user tests, PUPs covered 18 of those contexts. Evaluators predicted problems in nine contexts not discovered in user test. Note that the definition of a context needs further investigation. Other researchers have moved away from using problem counts and used a more qualitative approach with an analysis of types of usability issues and user-system misfits (Connell et al., 2004). Another step towards acknowledging the qualitative nature of the problem descriptions, instead of using merely the stringent concepts of Hits, Misses and False Alarms as is traditionally used, is to introduce Possible Hits (PH), which are less clear-cut, ambiguous, but nevertheless plausible matches between two methods, and Not Directly Observable consequences (NDO) (Connell et al., 2004). Therefore, false positives and NDOs are restricted to expert analysis and may be given less priority in revisions unless their frequencies or severities give rise to other actions. An empirical study of how developers prioritize usability problems to correct and whether there is a difference in problem revision strategies for two sets of usability problems being derived from

different usability evaluation methods should help shed light on this issue.

6. Concluding remark

The framework for comparing evaluation methods that we have described in this paper can be reused by other researchers because of its thorough structure. The study can be seen as a first one of this framework but subsequent studies may show that it needs to be improved. As a final phase of this framework, we will in this section refine a research agenda for comparing and contrasting evaluation methods.

While no conclusive claims about the two variables of interest – two sets of usability heuristics and two media of problem reporting – can be derived from the results of the present study, some implications for usability practitioners and researchers can be drawn. From the practical point of view, a web tool for capturing and recording usability problems in heuristic evaluation is recommended, especially when remote evaluation becomes increasingly popular, thanks to the relentless expansion of the Internet. Besides, easy access to and effective management of the data being captured in usability evaluation will enable collaborative efforts of practitioners and researchers distributed in different locations to work on common problems of interest. Further development of the tool creates opportunities for more intelligence in all aspects of evaluation, namely inspective, descriptive and analytical parts. In other words, the tool can improve the way we conduct the inspection of user interfaces, the way we report usability problems thus identified, and the way we consolidate a list of usability problems (i.e. eliminating duplicates as well as False Alarms and accurate severity rating), enabling practitioners to prioritize and correct urgent problems.

Similar to our two sets of heuristics, the literature has examined two different types of heuristics. One type is synthesized, (Connell and Hammond, 1999; Nielsen, 1994a), that is, a heuristic set is created bottom up from a larger set of heuristics or types of usability problems. Another type of heuristics is defined from abstract theories (Hornbæk and Frøkjær, 2004). The potential challenges that evaluators have with the former type is that, when researchers or practitioners find a common title for a synthesized heuristic, some details may be lost. With the second type, when the heuristics are described concretely, with one or two examples, evaluators may miss problems if they are unable to understand the abstract description.

Our evaluators appreciated the help they received through examples in the tool. "Learning by Examples" is a well-researched topic in cognitive psychology (Renkl and Atkinson, 2003). A caveat needs to be made that examples themselves should be of good quality; otherwise, they would impede rather than facilitate learning. Moreover, examples can rigidify how learners interpret heuristics, i.e. lower their free creative responses. Zhang's perspective based evaluation (Zhang et al., 1999) asks the evaluator to view the human computer interaction with a certain perspective, i.e. user type. Concrete examples, e.g. of novice and expert users are given. This method has resulted in 30% increase of usability problems over heuristics evaluation for 3 evaluators.

One recommendation that comes out of the work presented in this paper is that better training schemes need to be devised. More good concrete examples need to be shown, but also help with understanding the abstract meaning behind the heuristics, which will give evaluators enough freedom to identify problems not explicitly listed in the training material. Depending on the knowledge and experience gaps being identified, training should be adapted and personalized to specific profiles of individual evaluators. To help devise those examples and abstract tools, the following research question is posed:

- How can usability heuristics help evaluators identify problems? What is the cognitive mechanism underpinning heuristic evaluation? Has the heuristic or principle named really guided the evaluator to uncover a usability problem or the evaluator named it to justify her/his behaviour? Indeed, Cockton and Woolrych (2001) attempted to check the accuracy of heuristics that their evaluators attributed to usability problems. The corresponding metric was coined as *appropriateness*. Accordingly, appropriate heuristic applications can be determined by correspondence between predicted difficulties and applicability criteria as stated in a HE training manual (Lavery et al., 1996). Nonetheless, we assume that the reliability of such accuracy checks varies with assessors' level of expertise, both in the heuristics and application domain.

One of the reasons that Gerhardt-Powals' set of heuristics did not yield better outcomes than Nielsen's may be that it did not fit the application of a brokerage system. Furthermore, with the high complexity of an application but limited resource, evaluation thus needs to be scoped and goal-oriented. Somervell's and McCrickard's (2005) work focuses on the creation of so-called 'critical parameter-based' heuristics, based on critical parameters proposed by Newman and Taylor (1999). For the class of large-scale display systems Somervell and McCrickard identified three critical parameters, namely Interruption, Reaction and Comprehension. Clearly, the mapping of heuristics to the system's characteristics is important. But Somervell and McCrickard (2005) emphasize that mapping is not at the level of an individual system but a class of systems. This assertion aligns with Nielsen's recommendation *"Furthermore, it is possible to develop category-specific heuristics that apply to a specific class of products as a supplement to the general heuristics"* (Nielsen, 1994b, p. 29). He further suggests performing competitive analysis and user testing to create abstract categories of specific heuristics, but Somervell and McCrickard (2005) use claims-analysis as a basis for heuristic creation. A registration tool may help collect

the most commonly found faults in this class of applications or design features, depending on the granularity of the subject of evaluation. For brokerage systems, critical parameters could be Relevance (search results), Ease of upload/download (learning objects) and Security (intellectual property right). To further develop this issue we formulate another research question:

- How should a set of usability heuristics be so selected that they can best fit the context of the application domain, the goal of the evaluation and available resources? Is it possible to develop meta-guidelines to address this context fitness?

One reason for the low thoroughness of the two heuristic sets may be the large set of features or task scenarios that are described to the evaluator. Instead of inspecting an application for a few hours, a more iterative approach in a series of inspection sessions where a set of features or tasks are inspected each time may be more effective (Molich et al., 2004; Nielsen, 1994b). Inspection work is very tedious and tiring, and it could be that at the end of the session evaluators stopped or showed decreased performance because of fatigue. An iterative inspection session could reveal more differences between the two heuristic sets lists. Such an iterative approach may also give rise to a debriefing session in between iterations. Thus, evaluators can discuss problems discovered and, hence, raise their understanding of the heuristics or the domain. A third research question we pose is:

- Are evaluation results not only dependent on number of evaluators but also on the number and duration of iterative inspection sessions?

In summary, there remain challenges for usability practitioners and researchers to overcome. To cope with the problem of generalizability and transferability across contexts, extensive collaboration within the usability community to conduct multi-site experiments and to support exchange of ideas and experiences is deemed essential.

References

Andre, T.S., Hartson, R.H., Belz, S.M., McCreary, F.A., 2001. The user action framework: a reliable foundation for usability engineering support tools. International Journal of Human-Computer Studies 54, 107–136.

Barnum, C.R., 2003. What's in a number. Available from : <http://www.stcsig.org/usability/newsletter/0301-number.html>, Accessed 2006, Society for Technical Communication.

Bastien, J.M.C., Scapin, D., Leulier, C., 1999. The Ergonomic Criteria and the ISO 9241-10 Dialogue Principles: a comparison in an evaluation task. Interacting with Computers 11, 299–322.

Blackmon, M.H., Kitajima, M., Polson, P.G., 2005. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In: Conference on Human Factors in Computing Systems archive Proceedings of the SIGCHI Conference on Human factors in Computing Systems, ACM, Portland, Oregon, USA, pp. 31–40.

Brooke, J., 1996. SUS: A 'quick and dirty' usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (Eds.), Usability Evaluation in Industry. Taylor & Francis, London, pp. 189–194.

Chattratichart, J., Brodie, J., 2004. Applying user testing data to UEM performance metrics. In: CHI 2004, ACM, Vienna, Austria, pp. 1119–1122.

Cockton, G., Lavery, D., 1999. A framework for usability problem extraction. In: Sasse, A., Johnson C. (Eds.), INTERACT 1999, pp. 347–355.

Cockton, G., Woolrych, A., 2001. Understanding inspection methods: lessons from an assessment of heuristic evaluation. In: Blandford, A., Vanderdonckt, J., Gray, P.D. (Eds.), People and Computers XV. Springer-Verlag, Lille, France, pp. 171–182.

Cockton, G., Lavery, D., Woolrych, A., 2003a. Inspection-based evaluation. In: Jacko, J.A., Sears, A. (Eds.), The Human-Computer Interaction Handbook, Lawrence Erlbaum Associates, NJ.

Cockton, G., Woolrych, A., Hall, L., Hindmarch, M., 2003b. Changing analysts' tunes: the surprising impact of a new instrument for usability inspection method assessment. In: Palangue, P., Johnson, P., O'Neill, E. (Eds.), HCI 2003. Springer-Verlag, Bath, pp. 145–162.

Cockton, G., Woolrych, A., Hindmarch, M., 2004. Reconditioned merchandise: extended structured report formats in usability inspection. In: CHI'2004, ACM, Vienna, Austria, pp. 1433–1436.

Connell, I.W., Hammond, N.V., 1999. Comparing usability evaluation principles with heuristics. In: Sasse, A., Johnson C. (Eds.), Proceedings of the 7th IFIP international conference on Human-computer Interaction, INTERACT'99, IOS Press, Edinburgh.

Connell, I., Blandford, A., Green, T., 2004. CASSM and cognitive walkthrough: usability issues with ticket vending machines. Behaviour and Information Technology 23, 307–320.

Cook, D.J., Bailey, B.P., 2005. Designers' use of paper and the implications for informal tools. In: OZCHI 2005, vol. 122, ACM, Canberra, Australia, pp. 1–10.

Desurvire, H.W., Kondziela, J.M., Atwood, M.E., 1992. What is gained and lost using evaluation methods other than empirical testing. In: Monk, A., Diaper, D., Harrison M.D. (Eds.), HCI, Proceedings of the Conference on People and Computers VII, pp. 89–102.

Dillon, A., Watson, C., 1996. User analysis HCI – the historical lessons from individual differences research. International Journal of Human-Computer Studies 45, 619–638.

Doubleday, A., Ryan, M., Springett, M., Sutcliffe, A., 1997. A comparison of usability techniques for evaluating design. In: DIS'97, ACM Press, Amsterdam, pp. 101–110.

Faulkner, L., 2003. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments, and Computers 35, 379–383.

Folmer, E., Bosch, J., 2004. Architecting for usability: a survey. The Journal of Systems and Software 70, 61–78.

Frøkjær, E., Lárusdóttir, M.K., 1999. Predictions of usability: comparing method combinations. In: Managing Information Technology Resources in Organizations in the Next Millennium, Idea group publishing.

Gerhardt-Powals, J., 1996. Cognitive engineering principles for enhancing human-computer performance. International Journal of Human-Computer Interaction 8, 189–211.

Gladwell, M., 2002. The social life of paper. In: The New Yorker, The New Yorker Magazine , Newyork, NY, pp. 92–96.

Gray, W.D., Salzman, M.C., 1998. Damaged merchandise? Human-Computer Interaction 13, 203–262.

Hartson, H.R., Andre, T.S., Williges, R.C., 2001. Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction 13, 373–410.

Hertzum, M., Jacobsen, N.E., 2001. The evaluator effect: a chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction 13, 421–443.

Holzinger, A., 2005. Usability engineering methods for software developers. Communication of the ACM 48, 71–74.

Hornbæk, K., Frøkjær, E., 2004. Usability inspection by metaphors of human thinking compared to heuristic evaluation. International Journal of Human-Computer Interaction 17, 357–374.

Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M., 1991. User interface evaluation in the real world: a comparison of four techniques. In: ACM CHI'91, New Orleans, LA, pp. 119–124.

John, B.E., Mashyna, M.E., 1997. Evaluating a multimedia authoring tool. Journal of the American Society for Information Science 48, 1004–1022.

Karat, C., Campbell, R., Fiegel, T., 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In: ACM CHI'92 Conference, Monterey, California, pp. 397–404.

Lavery, D., Cockton, G., Atkinson, M.P., Heuristic Evaluation. Usability Evaluation Materials, <http://www.dcs.gla.ac.uk/asp/materials/HE_1.0/>, 1996.

Law, E.L.-C., Hvannberg, E.T., 2004a. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In: NordiCHI 2004, Tampere, Finland, pp. 241–250.

Law, E.L.-C., Hvannberg, E.T., 2004b. Analysis of the combinatorial user effect of international usability tests. In: CHI 2004, ACM, Vienna, Austria, pp. 9–16.

Lewis, J.R., 1991. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. In: ACM SIGCHI Bulletin, vol. 23, pp. 78–81.

Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M., 2003. Heuristic evaluation of ambient displays. In: CHI 2003, ACM Press, Florida, USA., pp. 169–176.

Molich, R., Jeffries, R., 2003. Comparative Expert Reviews. In: CHI 2003: New Horizons, ACM, Ft. Lauderdale, Florida, USA, pp. 1060–1061.

Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B., 2004. Comparative usability evaluation. Behaviour and Information Technology 23, 65–74.

Newman, W., Taylor, A., 1999. Towards a methodology employing critical parameters to deliver performance improvements in interactive systems. In: Sasse, M.A., Tauber M. (Eds.), INTERACT'99, 7th IFIP TC.13 International Conference on Human-Computer Interaction, Edinburgh, Scotland, pp. 605–612.

Nielsen, J., 1993. Usability Engineering. Academic Press, New York.

Nielsen, J., 1994a. Enhancing the explanatory power of usability heuristics. In: CHI'94, ACM, Boston, Massachusetts USA, pp. 152–158.

Nielsen, J., 1994b. Heuristic evaluation. In: Nielsen, J., Mack, R.L. (Eds.), Usability Inspection Methods. John Wiley & Sons, pp. 25–62.

Nielsen, J., 2000. Why you only need to test with 5 users? Available from: <http://www.useit.com/alertbox/20000319.html>, Accessed 2005.

Nielsen, J., Molich, R., 1990. Heuristic evaluation of user interfaces. In: Proceedings ACM CHI'90 Conference, ACM, Seattle, WA, pp. 249–256.

Norman, D.A., 1986. Cognitive engineering. In: Norman, D.A., Draper, S.W. (Eds.), User Centered System Design: New Perspectives on Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 31–61.

Renkl, A., Atkinson, R.R., 2003. Structuring the transition from example study to problem solving in cognitive skills acquisition: a cognitive load perspective. Educational Psychologists 38, 15–22.

Rieman, J., Davies, S., Hair, D.C., Esemplare, M., Polson, P., Lewis, C., 1991. An automated cognitive walkthrough. In: Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, ACM, New Orleans, Louisiana, United States, pp. 427–428.

Rosenbaum, S., Ronn, J.A., Humburg, J., 2000. A Toolkit for Strategic Usability: Results from Workshops, Panels, and Surveys. In: Turner, T., Szwillus, G., Czerwinski, M., Paterno F. (Eds.), CHI'2000, ACM, Hague, Amsterdam, pp. 337–344.

Somervell, J., McCrickard, D.S., 2005. Better discount evaluation: illustrating how critical parameters support heuristic creation. Interacting with Computers 17, 592–612.

Somervell, J., Wahid, S., McCrickard, D.S., 2003. Usability heuristics for large screen information exhibits. In: Rauterberg, M., Menozzi, M., Wesson J. (Eds.), INTERACT 2003, Zurich, Switzerland., pp. 904–907.

Spool, J., Schroeder, W., 2001. Testing web sites: five users is nowhere near enough. In: CHI '01 Extended Abstracts on Human factors in Computing Cystems, Seattle, Washington, pp. 285–286.

Virzi, R.A., 1992. Refining the test phase of usability evaluation: How many subjects is enough? Human Factors 34, 457–468.

Woolrych, A., Cockton, G., 2001. Why and when five test users aren't enough. In: Vanderdonckt, J., Blandford, A., Derycke, A. (Eds.), IHM-HCI, vol. 2. Toulouse, France, pp. 105–108.

Woolrych, A., Cockton, G., Hindmarch, M., 2004. Falsification testing for usability inspection method assessment. In: Fincher, S., Markopoulos, P., Moore, D., Ruddle R. (Eds.), HCI, BCS, Bath.

Zhang, Z., Basili, V., Shneiderman, B., 1999. Perspective-based usability inspection: an empirical validation of efficacy. Empirical Software Engineering 4 (1), 43–69.