A Longitudinal Study of Usability in Health Care

Does Time Heal?

Kjeldskov, Jesper; Skov, Mikael B.; Stage, Jan

Published in:
International Journal of Medical Informatics

DOI (link to publication from Publisher):
10.1016/j.ijmedinf.2008.07.008

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

A longitudinal study of usability in health care: Does time heal?

Jesper Kjeldskov, Mikael B. Skov, Jan Stage*

Aalborg University, Department of Computer Science, Selma Lagerloefs Vej 300, DK-9220 Aalborg East, Denmark

ARTICLE INFO

Article history:
Received 31 October 2007
Received in revised form 6 July 2008
Accepted 9 July 2008

Keywords:
Electronic patient records
Usability
Longitudinal study
Experts and novice users

ABSTRACT

We report from a longitudinal laboratory-based usability evaluation of a health care information system. The purpose of the study was to inquire into the nature of usability problems experienced by novice and expert users, and to see to what extend usability problems of a health care information system may or may not disappear over time, as the nurses get more familiar with it—if time heals poor design? As our method for studying this, we conducted a longitudinal study with two key studies. A usability evaluation was conducted with novice users when an electronic patient record system was being deployed in a large hospital. After the nurses had used the system in their daily work for 15 months, we repeated the evaluation. Our results show that time does not heal. Although some problems were not experiences as severe, they still remained after 1 year of extensive use. On the basis of our findings, we discuss implications for evaluating usability in health care.

1. Introduction

Usability evaluations are increasingly applied to assess the quality of interactive software systems. Usability has been defined as consisting of three aspects: efficiency, effectiveness and satisfaction and is often also measured on the basis of identified of usability problems [11,15,16]. Most mainstream approaches to usability evaluation involve "prospective users" thinking aloud while using the system [7,16,18]. According to mainstream guidelines, there is a considerable difference between involving so-called novice or expert users because these users may have different levels of experience with the system being evaluated. However, the consequence of involving novice or expert users as test subjects when evaluating a system's usability is still being debated (see for example Ref. [17]) and several comparative studies are being reported (see for example Refs. [3,9,17,19]). The questions remaining unanswered are many. To name a few, how are the results produced from an evaluation with novice users different from the results produced from an evaluation with experts? How is efficiency, effectiveness, experienced usability problems and subjective satisfaction or workload different from novice users to experts? To what extend does the time spent acquiring expertise with a system help users overcome its usability problems? These questions are highly relevant when evaluating the usability of information systems in health care.

Inspired by Nielsen [14], the purpose of the study reported in this paper is to inquire into nurses' experience of a health care information system over time as they develop system expertise. The key question is how the nurses' experience of the system's usability changes when they transform from being novices to being experts. Do usability problems disappear when users get more familiar with a system? Does time heal poor design? Addressing these overall questions, we report from an experiment comparing the experienced usability of an electronic patient record system when it was introduced into a large hospital to the experienced usability after 1 year of extensive use. The results of this experiment

* *Corresponding author*. Tel.: +45 9940 8921.
E-mail addresses: jesper@cs.aau.dk (J. Kjeldskov), dubois@cs.aau.dk (M.B. Skov), jans@cs.aau.dk (J. Stage).
1386-5056/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2008.07.008

are presented in detail and discussed as a basis for advising evaluators on selection of test subjects and design of task assignments when preparing a usability evaluation within the health care domain.

1.1. *What was known before the study?*

The study was based on the following knowledge and assumptions at the time.

- There is a considerable difference between involving novice or expert users because these users may have different levels of experience with the system being evaluated. However, the consequence of involving novice or expert users as test subjects when evaluating a system's usability is unclear.
- There is an assumption that usability problems may "disappear" over time as users transition from being novices to being experts. However, it is unclear if this assumption is justified and to what extend time spent acquiring expertise with computer systems help users overcome its usability problems.

2. Evaluating with novice and expert users

The human–computer interaction (HCI) literature generally discusses the importance of using "appropriate test subjects" when carrying out a usability evaluation. Typically, it is pointed out, that it is vital to choose participants that are representative of the intended target user community with respect to parameters such as their demographic profile (sex, age, education, profession, etc.), and their level of experience (for example if they are novices or experts) [6,16,18]. In relation to the level of user expertise, Nielsen [15] propose that there are (at least) three different dimensions to consider:

(1) The user's knowledge about the domain (ignorant versus knowledgeable).
(2) The user's experience with computers in general (minimal versus extensive).
(3) The user's experience with the system being evaluated (novices versus experts).

In relation to system experience, the discussion of when and why to choose test subjects with high or low level of experience is still ongoing. Some systems are only intended to be used infrequently by first-time users, such as many web-based systems, installation programs, etc., and should thus support novices by being quick and easy to learn. Other systems, such as airline booking systems, advanced industrial control systems, and many systems within the health care domain are designed for more frequent use and for highly experienced users. These may take longer time to learn to use but should, in the long run, support expert users by being highly effective. When evaluating such systems it is often intended to have test subjects that reflect the expected profile of the end users. However, in reality it is often difficult and sometimes not even possible to make such a simplistic differentiation between novice and expert users [15]. In real life, users often do not acquire expert skills in all parts of a system regardless of how much they use it because most systems are often very complex and offer a wide range of features that are not frequently used. Thus even highly experienced users of a system may still be novices in respect to some parts of it. Likewise, novice users of a system may have a high enough level of expertise with, for example, the use domain or computers in general to be able to understand and operate even very complex new systems if they are designed properly. Also, it is commonly known that test subjects may feel under considerable pressure during a usability evaluation because they feel that *they* are being assessed and not the system [16,18]. For novice users, this feeling of insecurity may be higher than for experts because they are not familiar with the system, and more efforts may consequently be required for making the test subject feel comfortable with the situation [18]. On the other hand, when testing with experts, some usability problems may not appear because these users have developed workarounds to compensate for poor design, yet the problems are still there. A final issue is access to test subjects. While it is typically not a problem to find novice users, it can sometimes be difficult to gain access to a large number of system experts, especially if the system is still under development or has not yet been deployed in the target organization.

Several experiments have inquired into the difference between novices and experts. In information retrieval, it has been observed that novice users often perform poorly [1]. An empirical study of information retrieval through search in a database compared the performance of novices and experts. Though there were no significant differences in the accuracy with which tasks were solved, the expert users performed significantly faster than the novices [5]. In a usability evaluation of a nursing assessment system, novices experienced severe usability problems that were not experienced by the experts. The novice users could not complete the tasks without going back to the patient for more information, and had difficulties locating where information should be entered into the system. The experts, on the other hand, could complete the tasks and had learned to use the system as a checklist for collecting the necessary information [4].

The empirical studies mentioned above all share the characteristic that experiments with novices and experts are conducted at the same time. Thus these experiments rely on a *classification* of different people as experts and novices. Such a classification is not without problems [2]. Our aim with the study reported in this paper has been to examine the difference between novice and expert user performance within the health care domain but based on a longitudinal study involving the *same users* in both evaluations. We have focused on the following research questions:

- RQ1: To what extent is the effectiveness and efficiency of using an electronic patient record (EPR) system different from novices to experts and is this measure identical for different tasks?
- RQ2: Which usability problems of an EPR system are experienced by novices and by experts: which problems are the same, and is there a difference in the severity of the problems experienced?

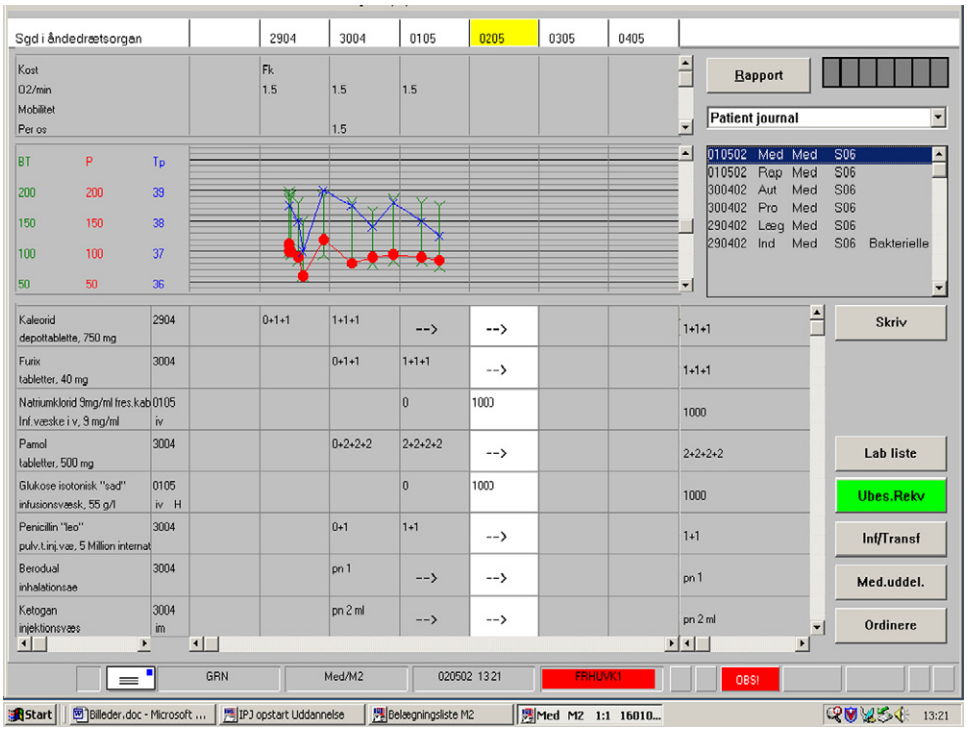

Fig. 1 – The status window of the EPR system.

- RQ3: How do novices and experts perceive the workload involved when solving work tasks using the system?

The first question reflects two of the fundamental aspects of usability. Although they seem related, it has been shown empirically, that it is necessary to consider both, as they are not always correlated [7]. The next question focuses on the usability problems experienced by novices and experts both in terms of the problems and their severity. Finally, the third question deals with the workload. As emphasized above, novice users tend to find usability evaluations very demanding. The third research question aims to provide a more firm foundation for that observation.

3. Electronic patient record usability

Between 2002 and 2003 we undertook a longitudinal empirical study of novice and expert users' experience of the usability of an electronic patient record (EPR) system for a large regional hospital in Denmark (IBM IPJ 2.3, Fig. 1). The basic design of the study was to conduct two usability evaluations of the same system with the same users. The first evaluation was conducted in May 2002 when the EPR system was being deployed at the hospital. The second evaluation was done in August 2003 when the users had used the system in their daily work for more than a year.

A key part of the system's use domain is the hospital wards. The nurses in each ward and the medical doctors use patient records to access and register information about their patients. They also use it to get an overview of the patients that are in a ward. Through the patient record, they can see the state, diagnosis, treatment, and medication of each individual patient. The nurses use the patient record in three different situations:

(1) Monitoring how the state of a patient develops.
(2) Daily treatment of a patient.
(3) Emergency situations.

The monitoring typically involves measurement of values, for example blood pressure and temperature. These values are usually measured at the patient's bed and typed in later. The daily treatment of patients can be described as structured problem solving. A nurse will observe a problem with a patient, for example that the temperature is high. She will then make a note about this and propose an action to be taken. This action is subsequently evaluated after some time. All steps are documented in treatment notes. In addition, the patient record provides a basis for coordination between nurses. For example, a nurse coming on duty will look through the list of patients to get an overview of their status and to check the most recent treatment notes to see what treatment has been carried out and what treatment is pending.

Medical doctors and nurses have developed the traditional paper-based patient record as a manual document style over a long period of time. The aim of the electronic record is to computerize that manual document. An electronic patient record is confronted with all the classical problems of creating a database that is shared across a complex organization and designing an interface that is both easy and effective to use. In addition, a hospital has many different groups of employees who may record and interpret data differently. The advantages of electronic patient records are also classical. The

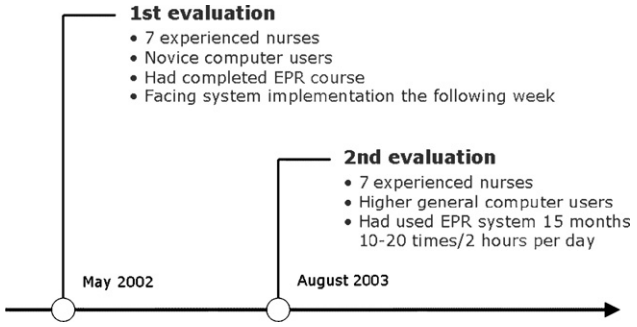

Fig. 2 – Timeline of the longitudinal study.

primary one is that data will be accessible to all personnel at all times whereas paper-based patient records usually follow the patient physically and is only accessible at one physical location at a time. Electronic patient records also potentially make overall processing of information about large groups of patients much easier.

The system used in our study was IBM's electronic patient record system IPJ 2.3 (Fig. 1). To facilitate our study, IBM personnel installed the IPJ 2.3 system in our usability laboratory and configured it to match the system used at the hospital in collaboration with two nurses dealing with the training and deployment of the system at the hospital. The nurses also created fictive but realistic patient data for the test setup.

4. Method

Below, we describe the method of the two usability evaluations of the EPR system as outlined in Fig. 2.

4.1. *The novice and expert users*

The first usability evaluation involved seven trained nurses from the same hospital. Prior to this evaluation, they had all attended a course on the IPJ system, and they were just starting to use the system in their daily work. All seven nurses were women, aged between 31 and 54 years, their experience as nurses varied between 2 and 31 years. Before the first evaluation they had received between 14 and 30 h of training in the EPR system. They characterized themselves as novices in relation to the EPR system and IT in general.

The purpose of the second evaluation was to facilitate a longitudinal study of the usability of the system after 1 year of use. In order to avoid the source of error that originates from individual differences between randomly selected test subjects we used the same seven participants in both evaluations. Before the second evaluation, all the nurses had used the system in their daily work for about 15 months. They indicated that they on average used the system 10–20 times a day, amounting to a total time of use of about 2 h per day. Therefore, we now characterized them as experts.

4.2. *Preparations*

In preparation for the evaluations, we visited the hospital and had a number of meetings and discussions with the two nurses who trained the personnel in the EPR system and dealt with the deployment of it. The purpose was to understand the work at the hospital wards related to patient record use and to get an overview of the system. Based on this we made a number of use scenarios for the system in collaboration with the nurses who were responsible for the deployment of the system.

4.3. *Tasks*

The purpose of the usability evaluations was to inquire into the usability of the EPR system for supporting nurses in solving typical work tasks. Based on our scenarios, we designed seven tasks, including a number of subtasks, centred on the core purpose of the system such as retrieving information about patients, registering information about treatments, making notes, and entering measurements. The tasks were developed in collaboration with the two nurses dealing with the implementation of the EPR system at the hospital. The exact same tasks were used in both evaluations. The tasks conformed to general guidelines for usability evaluation tasks briefly stating: (1) the overall scenario, for example, "it is 2nd May, you are the duty officer and have just come back to work after a long weekend. You are about to plan work tasks for the day for yourself, an assistant and two nurse students", (2) the specific background for the task with suitable detail, for example, "there are two new patients in the ward, Agnes Winther who needs physical assistance, and Casper Hansen who is on diet", and (3) the tasks themselves, for example "plan today's work tasks for the patients on the ward based on the information available in the system". For each scenario here was 2–3 specific subtasks to carry out. The seven tasks, and individual subtasks, resembled routine work with a few prompts of functionality used only occasionally. The tasks followed on from each other in a logical sequence resembling progress through a normal workday. The seven tasks did not cover all of the system's functionality. Like other electronic patient record systems, the functionality of the IPJ system is quite comprehensive, and engaging with all parts of it would require at least a whole day per test subject, which is not feasible for a usability evaluation. Instead, the tasks covered selected parts of the system functionality, which had been identified by the nurses as most important and most frequently to be used.

4.4. *Test procedure*

The test sessions were based on the "think-aloud" protocol as described by Rubin [18] and Nielsen [14] where the test subjects solve a series of tasks while thinking-out loud, describing their actions, how they perceive the system etc. In both evaluations, the seven test sessions were conducted over 2 days. The order of the nurses was random. Each nurse used the system to solve the seven tasks. This lasted approximately 45 min. If a test subject had problems with a task and could not continue on her own, the test monitor provided her with help to find a solution. If a test subject was completely unable to solve a task, the test monitor asked her to go on to the next one. One of the authors acted as test monitor throughout all 14 test sessions.

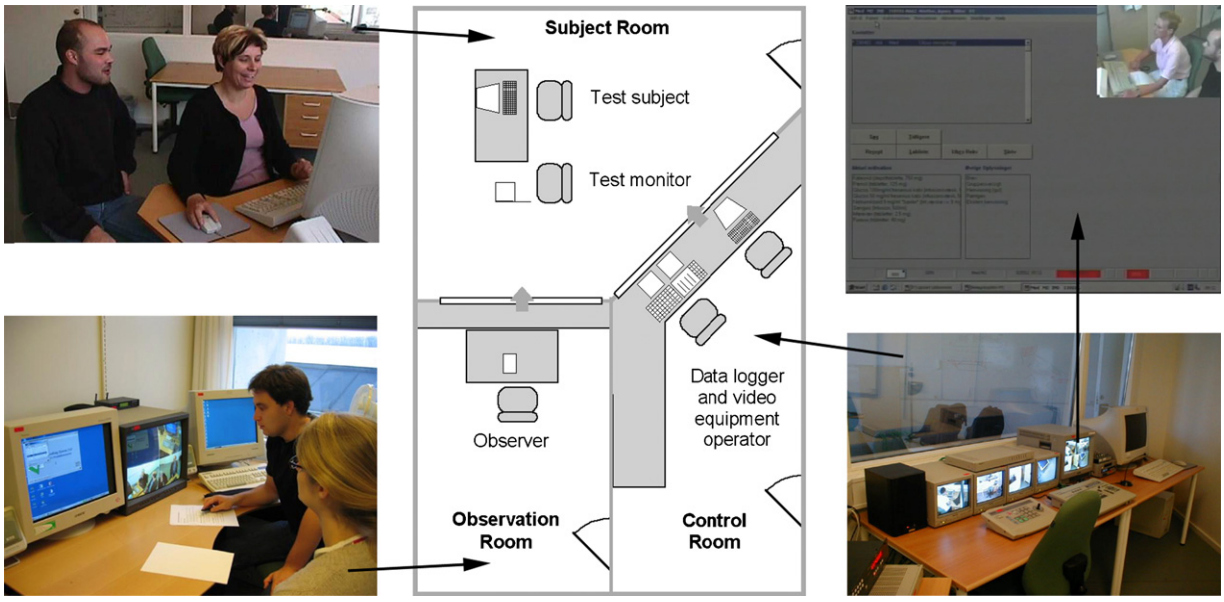

Fig. 3 – The test setup.

4.5. *Test setting*

All test sessions were conducted in a dedicated state-of-the-art usability laboratory at Aalborg University, Denmark (Fig. 3). For the EPR evaluation, the subject room was equipped with a desktop PC setup matching the hardware used at the hospital. The workload measurements were made in a separate room. The subject room could be monitored on video screens and through one-way mirrors from the adjacent observation and control rooms.

4.6. *Data collection*

All 14 test sessions were recorded on digital video. The video recording contained the PC screen with a small image of the test subject and test monitor inserted in the corner. The time spent on solving each task was measured from the video recordings. This measure is relevant for addressing RQ1. Immediately after each test, a workload measurement was made. This was based on the NASA task load index (TLX) technique. This measurement is intended to assess the user's subjective experience of the overall workload and the factors that contribute to it [8]. This measure was necessary for addressing RQ3. The two authors of this article who did not serve as test monitor, switched between conducting workload measurements and operating the laboratory equipment. Because of heavy time constraints on access to nurses, workload was only measured for four of the seven test subjects in the first evaluation. In the second evaluation we were able to measure workload for all seven participants.

4.7. *Data analysis*

The data analysis reported in this paper was conducted in August 2004, 1 year after the second evaluation. The two authors who did not serve as test monitor analysed all 14 videos. Each video was given a code that prevented the evaluator from identifying the year and test subject. The videos were assigned to the evaluators in a random and different order. The evaluators produced two individual lists of usability problems with a precise description. A usability problem was defined as a specific characteristic of the system that prevents task solving, frustrates the user, or is not understood by the user, as defined by Molich [13] and Nielsen [15]. Each evaluator also made a severity assessment for instance of a usability problem. The typical practice with severity is to make one general severity assessment for each problem expressed on a three-point scale, e.g., cosmetic, serious, and critical [13]. Yet this general severity assessment introduces a fundamental data analysis problem. Two users may experience the same problem very differently, and it is rarely clear how individual differences influence the general assessment. Moreover, we wanted to understand to what extent the severity changed from novices to experts. Therefore, we rated severity based on the extent to which it impacted the work process of each individual user. The severity ratings were necessary for addressing RQ2.

The individual problem lists from the two evaluators were merged into one overall list of usability problems. This was done in a negotiation process where the problems were considered one at a time until consensus had been reached. Out of the total number of 103 usability problems, 64 were identified by both evaluators, 17 only by evaluator 1, and 22 only by evaluator 2. The overlap between problems identified by the 2 evaluators suggests a low presence of the evaluator effect [10] and thus a high reliability of the merged list of problems. The resulting problem list was the basis for addressing RQ2. The evaluators also produced a 2–4 page log file for each of the 14 test sessions containing the exact times and descriptions of the users' interactions with the EPR system. The log file also describes whether the user solves each task, and to what extent the test monitor provides assistance. The extent

to which each task was solved and the test monitor interference was necessary for addressing RQ1.

5. Findings

5.1. *Effectiveness and efficiency (RQ1)*

Effectiveness reflects the accuracy and completeness of the subjects achieving certain goals and this includes indicators of quality of solution and error rates. In this experiment, we distinguish between completely and partially solved tasks. The mean numbers of solved tasks for the expert subjects were 6.29 (S.D. = 1.11) tasks and for the novice subjects 3.57 (S.D. = 1.27) tasks and a Wilcoxon signed rank test shows significant difference $z = 2.116$, $p = 0.034$. Thus, we found that the test subjects solved significantly more tasks as expert subjects than as novice subjects. The calculated standard deviations indicate high variance for the novice subjects; in fact the novice subjects on numbers of solved tasks ranged from three to six whereas the expert subjects ranged from five to seven. All expert subjects solved all seven tasks either completely or partially while only two novice subjects solved all tasks. This difference is strongly significant according to a Chi-square test $\chi^2[1] = 6.667$, $p = 0.0098$. Considering only completely solved tasks, four expert subjects failed to solve all seven tasks within the given time frame while all seven novice subjects failed to solve all tasks completely, but this difference is not significant $\chi^2[1] = 3.000$, $p = 0.0833$.

In conclusion, the expert users were more effective than the novices. The experts solved significantly more tasks and there was less variation than among the novices.

Efficiency reflects the relation between the accuracy and completeness of the subjects achieving certain goals and resources spent in achieving them. Indicators often include task completion time, which we use in this experiment. Despite the significant higher number of solved tasks, we found no significant differences in mean values for the total task completion times $z = 1.402$, $p = 0.161$. The assignments enfold important variances and the two simple data entry tasks were solved faster by the experts, but we found no significant differences for any of the individual tasks.

In conclusion, the experts were faster for simple data entry tasks, though not significantly faster, and on more complex tasks there were no major differences.

5.2. *Usability problems and severity (RQ2)*

We identified a total number of 103 usability problems. The top most of these were related to the three overall themes of: (1) complexity of information, (2) poor relation to work activities, and (3) lack of support for mobility [12]. The novices experienced 83 of these 103 usability problems whereas the expert subjects experienced 63 (Table 1). Attributing severity to the identified usability problems, the highest experienced severity for each problem is used. We found that the novices experienced 93% of the critical problems (25 of 27 problems) while the experts experienced 70% (19 of 27 problems). Similar distributions were identified for the serious problems where the novices experienced 80% of the identified problems compared 61% for the experts. Finally, minor differences were found for cosmetic problems: 65% for novices against 50% for experts.

Table 1 – Total numbers of identified usability problems for the novices and experts

	Novice ($N = 7$)	Expert ($N = 7$)	Total ($N = 14$)
Critical	25	19	27
Serious	45	34	56
Cosmetic	13	10	20
All	83	63	103

Table 2 – Mean numbers of identified usability problems for the two setups

	Novice ($N = 7$)	Expert ($N = 7$)	z	p
Critical	5.29 (1.50)	3.29 (1.98)	1.420	0.156
Serious	17.29 (3.09)	9.14 (2.97)	2.159	0.031
Cosmetic	8.86 (2.41)	11.43 (2.76)	−1.876	0.061
All	31.43 (4.93)	23.86 (4.49)	2.159	0.031

Table 2 outlines results on mean numbers of identified problems for novices and experts. We found that the novice subjects experienced significantly more problems than the experts according to a Wilcoxon signed rank test $z = 2.159$, $p = 0.031$. However, this difference is mainly a result of more serious problems $z = 2.159$, $p = 0.031$, whereas we found no significant differences for the critical problems $z = 1.420$, $p = 0.156$ or the cosmetic problems $z = 1.876$, $p = 0.061$.

Fig. 4 outlines problems unique to the novice subjects, problems unique to the expert subjects, and problems experienced by both novices and experts. Forty of the 103 identified problems were experienced by the novice subjects only and most of these problems concerned simple data entry tasks such as typing in values for patients. Forty-three of the 103 identified problems were experienced by both novice and expert subjects and they typically concerned advanced data entry or solving judgment questions. Twenty problems were identified for experts only. These mainly concern functionality and services that the novices did not use for solving the same tasks, for example work task lists, because they were not familiar with those parts of the system.

Discarding problems experienced only by 1 test subject (unique problems), we see that most of the usability problems were identified in both the novice sessions and expert sessions (40 of the 61). Furthermore, we see that the experts experienced only five non-unique problems not experienced by any novice subjects. None of these five problems were critical, and

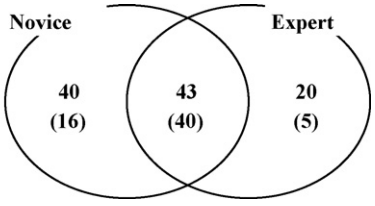

Fig. 4 – Distribution of the identified problems for the novices and experts. Numbers in parentheses show total numbers of problems subtracted unique problems.

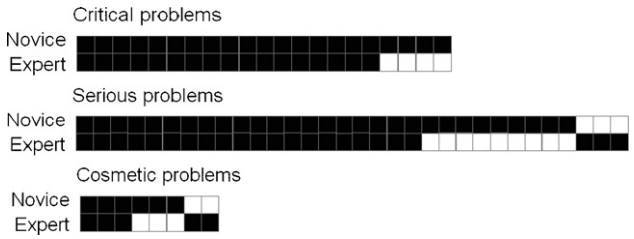

Fig. 5 – Distribution of usability problems identified by novices and experts. Each column represents a usability problem. A black square indicates that the respective user group identified a problem. A white square indicates that a problem was not identified by that user group.

accordingly, all critical non-unique problems were identified already in the first year of evaluation.

The distribution of usability problems experienced by more than one test subject is illustrated in Fig. 5. This figure shows that 17 critical problems experienced by the novices were still experienced after 1 year of use. Both novices and experts experienced more than half of the serious problems, while 9 serious problems were only experienced by the novices. The expert users, on the other hand, only experienced 3 serious problems not also experienced by the novices. In relation to the cosmetic problems, less than half were experienced by both novices and experts. Three cosmetic problems were experienced only by novices and 2 only by experts.

In conclusion, there was a huge overlap of both critical and serious usability problems experience by novices and experts. Some problems disappeared over time, but far from all of them. At the same time, new serious and cosmetic problems appeared because more parts of the system were being explored.

Based on our instrumentation for problem identification and categorization, we classified problems according to how the individual test subjects experienced the problems. Thus, the same problem could be critical to one subject while cosmetic to another. Forty-three of the 103 usability problems were experienced by both the novices and the experts. Attributing the severities values between 1 and 3 where 3 = critical, 2 = serious, and 1 = cosmetic problems, we can count the severity for each of the 43 problems. Considering the number of subjects experiencing the problems, each of the 43 problems was experienced on average by 3.61 (S.D. = 2.19) novice subjects and on average by 3.39 (S.D. = 2.01) expert subjects. But this difference is not significant according to a Wilcoxon signed rank test $z = 0.722$, $p = 0.470$. We further calculated the mean value for each of the 43 problems for the novices and experts. The mean value for novices was 1.91 (S.D. = 0.51) and the mean value for the experts was 1.55 (S.D. = 0.57) and this difference is significant $z = 3.963$, $p = 0.001$. Finally, we analyzed the problems experienced in both the first and second evaluation on worst-case for each year. Here we found that the problems on average had a value of 2.19 (S.D. = 0.59) whereas the experts on average experienced the problems to a mean value of 1.84 (0.75). This is significant according to a Wilcoxon signed rank test $z = 2.690$, $p = 0.007$.

In conclusion, a remarkably high number of problems was experienced both by novices and expert users. These problems were experienced significantly more severe for the novices, so the problems that remained became less severe after 1 year of use.

Table 3 – TLX-test values for the novice and expert subjects

	Novice ($N = 4$)	Expert ($N = 7$)
Mental	324 (109)	196 (97)
Physical	0 (0)	4 (8)
Temporal	61 (29)	29 (33)
Effort	306 (135)	135 (91)
Performance	138 (164)	164 (148)
Frustration	295 (94)	74 (52)
Sum	1124 (146)	602 (282)

5.3. *Task load index (RQ3)*

A NASA-TLX test was used to measure how the nurses experienced the testing situation. The NASA-TLX test is used to assess the subjective workload of people on six factors: effort, frustration, mental demand, performance, physical demand, and temporal demand. The subjects attribute the six factors with a value between 1 and 100 and the subjects assess the importance of these factors.

As evident in Table 3, the level of frustration and the total task load reduced dramatically, but the perceived effort and mental demands were still high. Most novice subjects expressed high frustration after the first evaluation. Specifically they found it frustrating that they were not able to solve the tasks properly and completely. In conclusion, the novices experienced frustration as significantly higher than the experts.

6. Implications for usability evaluations in health care

The implications for the choice of novice or expert users as test subjects are several. In relation to effectiveness, we found that the expert users completed significantly more tasks and had lower variance in task completion than the novices. This indicates that in situations where it is important for the software development process that every planned aspect of an expert system is evaluated, one should consider using experts rather than novices in order for the evaluation sessions not to be held up. As discussed in relation to efficiency, this does not necessarily influence task completion time.

In relation to identification of usability problems, we found a significant difference between the number of problems experienced by novices and experts. The implications of this finding are debatable. On one hand it can be stated that one should use novices because they enabled more problems to be found. On the other hand, it could be argued that the use of experts supported the elimination of noise from “false” usability problems (typically rated as cosmetic). Regardless, however, our results show that when evaluating a system designed for a specialized domain, such as health care, including users who are novices with the system but highly experienced with the use domain as test subjects can sup-

port the identification of as many critical and serious usability problems as when using system experts. This finding is important in situations where expert users may be a scarce resource for usability studies.

In relation to problem severity, we found a significant difference between the mean severity ratings for novices and experts, with the latter generally experiencing the usability problems of the system as less severe. The implications of this finding is primarily that when analyzing the data from a usability evaluation with novice users and making suggestions for subsequent response, designers should remember that even though time may not heal a system's usability problems, returning users will get familiar with the system, and that the cost associated with this learning may in some cases outweigh the costs of a redesign that may or may not be significantly better. This is especially important in relation to when responding to cosmetic usability problems.

Finally, in relation to the subjective experience of participating in an evaluation, we found that novices experienced significantly more mental workload and frustration than the experts. This is not surprising but stresses the fact that when testing with novices, the test monitor should be prepared to put more effort into making the test subjects feel comfortable with the situation as discussed in the novice–expert section above.

Summary points

What has the study added to the body of knowledge

On the basis of our findings, we propose these four summary points for usability evaluations in health care:

- Time does not heal. Although some problems were not experiences as severe, they still remained after one year of extensive use. Poor design remains poor.
- Expertise reduces experienced severity of usability problems. When testing with novice users, evaluators must take into consideration that some problems may not, in the long run, be as severe as it seems.
- Solve usability problems early. If usability problems do not disappear over time, we should get rid of them as soon as possible. There will always be novice users—new employees, temporary staff, etc.
- Evaluate with both novice and expert users and use their different experience of the system as a lens to get a more complete picture of a system's usability. They both represent a prospective user group.

7. Conclusions

This paper has reported from a longitudinal study in health care where we have compared the usability of an electronic patient record system as experienced by novice and expert users. The longitudinal study differed from other novice–expert studies because the seven test subjects remained the same throughout the study, and participated in the same test when they were new to the systems and after 1 year of extensive use. The usability of the system was measured in different ways. The first measure was effectiveness and efficiency. The expert users were more effective than the novices; they solved significantly more tasks and there was less variation than among the novices. However, we found no significant differences on task completion times for the individual tasks. The second measure was the number and severity of usability problems experienced. The novice subjects experienced significantly more critical and serious problems, whereas the experts experienced significantly more cosmetic problems. Thus there was a huge overlap of both critical and serious usability problems experienced by novices and experts. Some problems disappeared over time, but far from all of them. The identified problems were experienced significantly more severe for the novices, and some of the problems that remained after 1 year of use became less severe. At the same time, however, new serious and cosmetic problems appeared. The third measure was subjective workload. In relation to this, our study showed that frustration and the total task load was reduced dramatically, but that the perceived effort and mental demands were still high.

Some of the overall results confirm the outcome of other studies. The most striking results are that the expert users are not more efficient on complex tasks and that a remarkable number of serious and critical problems with the electronic patient record system still remained after one year of extensive use. Thus we conclude that time does not heal usability problems. Even though time allows people to learn strategies for overcoming a system's specific peculiarities, poor design remains poor.

The study reported in this paper also leaves several avenues for further research. One of the interesting questions is whether we can identify specific categories for the usability problems that remain and disappear respectively. In order to answer this question, more longitudinal studies must be conducted into the usability of interactive systems over time, focusing on qualitative characteristics of usability problems.

8. Limitations

There are a number of possible limitations to the study related to how our results can be generalized. Firstly, while the seven test subjects were carefully recruited to be representative of the workforce at the particular hospital in terms of their work expertise and demographic profile, they are not necessarily representative of the overall population beyond this domain. Although it is our assumption that time does generally not heal poor design, further studies will have to investigate if this is justified. Secondly, while it is our opinion that the system evaluated is, sadly, representative of many expert systems for complex data management in health care as well as in the corporate domain, it is not necessarily representative of all computerized systems. Again, further studies are needed to establish to what extend our results can be generalized to other systems.

Finally, it can be argued that there could be a learning effect between the two evaluations enabling the test subjects to perform better in the second evaluation because of re-exposure to

the tasks rather than higher expertise with the system. Given the huge daily exposure to the system over the year in between evaluations, and the general formulation of tasks to resemble real work activities at the hospital, we do not believe that using the same tasks again in the second evaluation impacted on the test subjects' performance as they all had to regularly solve those tasks in their daily work in between the two evaluations. Adding to this point, it is important to note that given our main point here is that *1 year of use did not eliminate the system's usability problems*, a presence of positive learning effects in the experimental design would only add to our conclusion that time does not heal the usability of poor design.

REFERENCES

[1] B. Allen, Cognitive abilities and information system usability, Information Processing and Management 30 (1994) 177–191.

[2] R.W. Bailey, R.W. Allan, P. Riello, Usability testing vs. heuristic evaluation: a head-to-head comparison, in: Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, HFES, 1992, pp. 409–413.

[3] R. Bednarik, M. Tukiainen, Effects of Display Blurring on the Behavior of Novices and Experts During Program Debugging. Extended Abstracts of CHI 2005, ACM, New York, 2005, pp. 1204–1207.

[4] P.Q. Bourie, J. Dresch, R.H. Chapman, Usability evaluation of an on-line nursing assessment, in: Proceedings of the AMIA Symposium, 1997.

[5] A. Dillon, M. Song. An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. Journal of Digital Information, 1 (1) (1997). http://journals.tdl.org/jodi/issue/view/1 (accessed July 5, 2008).

[6] A. Dix, J. Finlay, G.D. Abowd, R. Beale, Human–Computer Interaction, 3rd ed., Pearson Education Limited, Harlow, England, 2004.

[7] E. Frøkjær, M. Hertzum, K. Hornbæk, Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? CHI Letters 2 (1) (2000) 345–352.

[8] S.G. Hart, L.E. Staveland, Development of NASA-TLX (task load index): results of empirical and theoretical research, in: P.A. Hancock, N. Meshkati (Eds.), Human Mental Workload, Elsevier Science Publishers, Amsterdam, 1988, pp. 139–183.

[9] N. Ishii, K. Miwa, Interactive processes between mental and external operations in creative activity: a comparison of experts' and novices' performance, in: Proceedings of the Fourth conference on Creativity & cognition table of contents, New York, ACM, 2002, pp. 178–185.

[10] N.E. Jacobsen, M. Hertzum, B.E. John, The evaluator effect in usability tests, in: Proceedings of the CHI'98, ACM Press, New York, 1998.

[11] C.M. Karat, R. Campbell, T. Fiegel, Comparison of empirical testing and walkthrough methods in user interface evaluation, in: Proceedings of the CHI'92, ACM Press, New York, 1992, pp. 397–404.

[12] J. Kjeldskov, M.B. Skov, Exploring context-awareness for ubiquitous computing in the healthcare domain, Personal and Ubiquitous Computing 11 (7) (2007) 549–562.

[13] R. Molich, Usable Web Design, Ingeniøren|bøger, 2000 (in Danish).

[14] J. Nielsen (2000). Novice vs. Expert Users. Alertbox. http://www.useit.com/alertbox/20000206.html (February 6, 2000).

[15] J. Nielsen, Usability Engineering, Morgan Kaufmann, San Diego, 1993.

[16] J. Preece, Y. Rogers, H. Sharp, Interaction Design: Beyond Human–Computer Interaction, John Wiley & Sons, Inc., New York, 2002.

[17] J. Prümper, M. Frese, D. Zapf, F.C. Brodbeck, Errors in computerized office work: differences between novice and expert users, ACM SIGCHI Bulletin 23 (2) (1991) 63–66.

[18] J. Rubin, Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests, John Wiley & Sons, Inc., New York, 1994.

[19] H. Urokohara, K. Tanaka, K. Furuta, M. Honda, M. Kuros, NEM: "Novice Expert ratio Method" A Usability Evaluation Method to Generate a New Performance Measure. Extended Abstracts of CHI 2000, ACM, New York, 2000, pp. 185–186.