Proceedings of OZCHI 2005, Canberra, Australia. November 23 - 25, 2005.

DOES TIME HEAL? A LONGITUDINAL STUDY OF USABILITY

Jesper Kjeldskov, Mikael B. Skov and Jan Stage
Aalborg University, Department of Computer Science
Fredrik Bajers Vej 7, DK-9220 Aalborg East, Denmark
{jesper, dubois, jans}@cs.aau.dk

ABSTRACT

We report from a longitudinal laboratory-based usability evaluation of an interactive system. A usability evaluation was conducted with novice users when a large commercial electronic patient record system was being deployed in the use organization. After the users had used the system in their daily work for 15 months, same evaluation was conducted again. Our aim was to inquire into the nature of usability problems experienced by novice and expert users over time, and to see to what extends usability problems may or may not disappear over time, as users get more familiar with the system. On the basis of our two usability evaluations, we present key findings on the usability of the evaluated system as experienced by the two categories of users at these two different points in time. Based on our findings, we discuss implications for evaluating usability.

KEYWORDS: *Usability, Experts and Novices, Longitudinal, Electronic Patient Records*

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 (ISO 1997) and is often also measured on the basis of identified of usability problems (Karat et al. 1992, Nielsen 1993, Preece et al. 2002). Most mainstream approaches to usability evaluation involve "prospective users" thinking aloud while using the system (see e.g. Dix et al. 2004, Preece et al. 2002, Rubin 1994). 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 e.g. Nielsen 2000) and several comparative studies are being reported (see e.g. Bednarik and Tukiainen 2005, Ishii and Miwa 2002, Urokohara et al. 2000, Prümper et al. 1991,). 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 using a system heal its usability problems?

As proposed by Nielsen (2000), the purpose of this paper is to inquire into the difference between novice and expert users by studying users over time as they develop system expertise. The key question is how the user's experience of a system's usability changes when they transform from being novices to being experts – if usability problems really disappear over time when users get more familiar with a system. Addressing this overall question, we report from an experiment comparing the experienced usability of a system when it was introduced into a large organization to the experienced usability after one year of extensive use. The results of this experiment are presented in detail and discussed as a basis for advising evaluators on selection of test subjects and design of task assignments for the evaluation.

Additional copies are available at the ACM Digital Library (http://portal.acm.org/dl.cfm)
or ordered from the CHISIG secretary (secretary@chisig.org)
OZCHI 2005 Proceedings ISBN: 1-59593-222-4

2. EVALUATING WITH NOVICE AND EXPERT USERS

The Human-Computer Interaction 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 (e.g. if they are novices or experts) (Dix et al 2004, Preece et al. 2002, Rubin 1994). In relation to the level of user experience, Nielsen (1993) propose that there are (at least) three different dimensions of (1) the user's knowledge about the domain (ignorant versus knowledgeable), (2) the user's experience with computers in general (minimal versus extensive) and (3) the user's experience with the specific system being evaluated (novices versus experts).

Many guidelines for usability evaluation seem to rely on the assumption that there is a considerable difference between testing with novice or expert users of the system being evaluated (see e.g. Preece et al. 2002, Rubin 1994). However, it is not clear to what extent this assumption is justified and exactly how this difference influences the results of an evaluation. Thus, the discussion of whether to choose test subjects with high or low level of system 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 and advanced industrial control systems, 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 (Nielsen 1993). In real life, users often don't 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 (see e.g. Preece et al. 2002, Rubin 1994). 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 (Rubin 1994). 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. 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 enough 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 (Allen 1994). 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 (Dillon and Song 1997). 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 (Bourie et al 1997). The notion of mental model has been used to explain and inquire further into the differences between experts and novices. There are clear differences in this sense, as experts have mental models that are closer to the system's model (Kellogg and Breen 1987).

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 (Bailey et al. 1992). Our aim with the study reported in this paper has been to examine the difference between novice and expert performance

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 the system different from novices to experts and is this measure identical for different types of tasks?
- RQ2: Which usability problems are experienced by novices and by experts: which problems are the same and is there a difference in the severity of the problems that are experienced by both novices and experts?
- RQ3: How do novices and experts perceive the workload when solving work tasks that involve use of the system?

The first question reflects two of the fundamental aspects of usability (ISO 1997). Although they may seem related, it has been shown empirically, that it is necessary to consider both, as they are not correlated (Frøkjær et al. 2000). 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. With the third research question our aim is to provide a more firm foundation for that observation.

3. THE LONGITUDINAL STUDY

We undertook an empirical study of novice and expert users' experience of the usability of an interactive system. 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 system was being deployed in the user organization. The purpose of the second evaluation was to facilitate a longitudinal study of the usability of the system after one year of use. This evaluation was conducted in August 2003 when the users had used the system in their daily work for more than a year.

3.1. The Application Domain

The interactive system used in the study was an electronic patient record system for a hospital. A key part of the system's application domain is the hospital wards. The nurses in each ward and the medial 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, and (3) emergency situations. The monitoring typically involves measurement of values, e.g. 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, e.g. 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 the current status of the patients and to check the most recent treatment notes to see what treatment has been carried out and what treatment is pending.

3.2. Electronic Patient Records

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 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 (figure 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.

Figure 1: The status window of the IPJ system

3.3. 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 hours of training in the IPJ system. They characterized themselves as novices in relation to the IPJ system and IT in general.

The purpose of the second evaluation was to facilitate a longitudinal study of the usability of the system after one year of use. In order to avoid the source of error that originates from individual differences between randomly selected test subjects, we wanted to use 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 to 20 times a day, amounting to a total time of use of about 2 hours per day. Therefore, we now characterized them as experts.

3.4. The Two Usability Evaluations

Preparations: We visited the hospital and had a number of meetings and discussions with the two nurses who trained the personnel in the IPJ system and dealt with the deployment of it. The purpose was to understand the work in the hospital wards related to the patient record and to get an overview of the system. Based on this we made a number of scenarios of the use of the system in collaboration with the person who was responsible for the deployment of the system.

Tasks: The purpose of the usability evaluations was to inquire into the usability of the IPJ 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 IPJ system at the hospital. The exact same tasks were used in both evaluations.

Test Procedure: The test sessions were based on the think-aloud protocol as described by Rubin (1994) and Nielsen (1993). In both evaluations, the seven test sessions were conducted over two days. The order of the nurses was random. Each nurse used the system to solve the seven tasks. This lasted approximately 45 minutes. 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 of this article acted as test monitor throughout all 14 test sessions.

Test Setting: All test sessions were conducted in a dedicated state-of-the-art usability laboratory at Aalborg University, Denmark. We used a single test room, with a desktop PC setup matching the hardware used at the hospital. The workload measurements were made in a separate room.

Data Collection: All sixteen 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 (Hart and Staveland 1988, NASA). 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 4 of the 7 test subjects in the first evaluation. In the second evaluation workload was measured for all 7 participants.

Data Analysis: The data analysis was conducted in August 2004, one 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 (2000) and Nielsen (1993). In the individual problem lists, each evaluator also made a severity assessment for each test subject that experienced a usability problem. The typical practice with severity is to make one general severity assessment for each usability problem based on the description in the problem list. This assessment is expressed on a three-point scale, e.g. cosmetic, serious, and critical (Molich 2000). 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 novice to expert users, so we needed to know how each test subject experienced a usability problem. Therefore, we rated the severity for each test subject based on the extent to which it impacted the work process of that particular 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 two evaluators suggests a low presence of the evaluator effect (Jacobsen et al. 1998) 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 sixteen test sessions containing the exact times and descriptions of the users' interactions with the IPJ 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.

4. RESULTS

This section provides the quantitative results of the study. It is structured in accordance with the three research questions that were introduced above.

4.1. Effectiveness and Efficiency (RQ 1)

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 (SD=1.11) tasks and for the novice subjects 3.57 (SD=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 3 to 6 whereas the expert subjects ranged from 5 to 7. All expert subjects solved all seven tasks either completely or partially while only two novice subjects solved all tasks and this difference is strong 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 thus not significantly faster and we on more complex tasks there were no major differences.

4.2. Usability Problems and Severity (RQ 2)

Based on our analysis, we identified a total number of 103 usability problems. The novices experienced 83 of these 103 usability problems whereas the expert subjects experienced 63 of the 103 usability problems (this is shown in 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 the 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 outlines key results on mean numbers of identified problems for the 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. This difference is mainly a result of more identified 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.

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

As expressed in research question 2, we sought to explore differences and similarities in the problems identified by the two sets of subjects. Figure 2 outlines problems unique to the novice subjects, problems unique to the expert subjects, and problems experienced by both novices and experts. 40 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. 43 of the 103 identified problems were experienced by both novice and expert subjects and they typically concerned advanced data entry or solving judgment questions. 20 problems were identified for experts only. These mainly concern functionality and services that the novices did not use for solving the same tasks, e.g. work task lists.

Discarding unique problems from the distribution, we see that most of the usability problems (40 of the 61) were identified in both the novice sessions and expert sessions. Further, the experts experienced 5 non-

unique problems not experienced by any novice subjects and none of these 5 problems were critical. Accordingly, all critical non-unique problems were identified from the novice sessions.

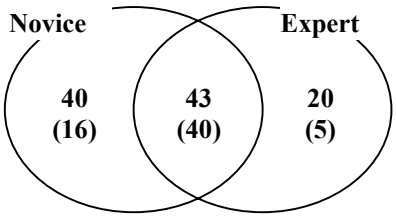

Figure 2: Distribution of the identified problems for the novices and experts. Numbers in parentheses show total numbers of problems subtracted unique problems.

The distribution of usability problems experienced by more than one test subject for the two groups is illustrated in figure 3 below. This figure shows the gaps and overlaps between problems experienced by novices and experts.

Figure. 3. Distribution of usability problems identified by novices and experts in the two studies. Each column represents a usability problem. A black square indicates that the respective user group identified a usability problem. A white square indicates that a problem was not identified by that user group but was found by the other user group.

As illustrated in figure 3, the novices experienced four critical problems that were not encountered one year later when the users had reached a higher level of expertise. It can be discussed whether these four problems were not really problems at all or if the expert users had just developed compensating workarounds for them. More importantly, one should notice that the remaining 17 critical problems experienced by the novices were still experienced after one year of use. Both novices and experts experienced more than half of the serious problems, while nine 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. 3 cosmetic problems were experienced only by the novices and 2 only by the 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.

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. 43 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 (SD=2.19) novice subjects and on average by 3.39 (SD=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 the problem for the novices was 1.91 (SD=0.51) and the mean value for the experts was 1.55 (SD=0.57) and this difference is significant $z=3.963$, $p=0.001$. Finally, we analysed 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 (SD=0.59) whereas the experts on average experienced the problems to a mean value of 1.84 (0.75) and this is significant according to a Wilcoxon signed rank test $z=2.690$, $p=0.007$.

In conclusion, a remarkably high number of problems were 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.

4.3. Task Load Index (RQ 3)

A NASA-TLX test was used to measure how the subjects 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.

Table 4. 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)

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. More of them 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.

5. IMPLICATIONS FOR USABILITY EVALUATION

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 (such as en electronic patent record) is evaluated, one should consider using experts rather than novices. As discussed in relation to efficiency, this does not necessarily influence task completion time in the parts of a system that has critical or serious usability problems.

In relation to the 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 identified. On the other hand, it could be argued that the use of experts supported the elimination of noise from “false” usability problems. Regardless of which of these points of views one may subscribe to, however, our results show that when evaluating a system designed for highly specialized domain, including users who are novices with the system but highly experienced with the use domain as test subjects can support 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 or non-existing resource.

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 actions to be taken, 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 well outweigh the costs of producing 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 may not be

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.

6. CONCLUSIONS

This paper has reported from a longitudinal study where we have compared the usability of an interactive system as novice and expert users experienced it. The longitudinal study differed from other novice-expert studies. The seven test subjects were the same, and they participated in the same test when they were new to the systems and after one 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 by the two groups. 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 experience by novices and experts. Some problems disappeared over time, but far from all of them. These problems were experienced significantly more severe for the novices, so the problems that remained became less severe. At the same time, new serious and cosmetic problems appeared.

The third measure was subjective workload. In relation to this, our study showed that the level of 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 substantiate 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 still remain after one year of extensive use. Thus we note that while time seems to mend some usability problems and reduce the severity of others by allowing people to learn strategies for overcoming a system's specific peculiarities, it far from heals all critical and serious usability problems.

On the basis of our findings, we have discussed a number of implications for evaluating usability contributing to the discussion of when and why to include novice or expert users of the system to be evaluated. 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 respectively disappear. 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.

7. REFERENCES

Allen, B. (1994). Cognitive Abilities and Information System Usability. *Information Processing and Management, 30,* 177-191.

Bailey, R.W., Allan, R.W., Riello, P. (1992). Usability Testing vs. Heuristic Evaluation: A Head-to-Head Comparison. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, HFES,* (409-413).

Bednarik, R. and Tukiainen, M. (2005). Effects of display blurring on the behavior of novices and experts during program debugging. In *Extended Abstracts of CHI 2005*, pp 1204 - 1207, New York: ACM

Bourie, P. Q., Dresch, J., and Chapman, R. H. (1997). Usability Evaluation of an On-line Nursing Assessment. In *Proceedings of AMIA Symposium.*

Dillon, A. and Song, M. (1997). 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).*

Dix, A., Finlay, J., Abowd, G.D. and Beale R. (2004). Human-Computer Interaction (third edition). Harlow, England: Pearson Education Limited.

Frøkjær, E., Hertzum, M. and Hornbæk, K. (2000) Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? *CHI Letters, 2(1).*345-352.

Hart, S.G., Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, P. A. and Meshkati, N. (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier Science Publishers.

Ishii, N. and Miwa, K. (2002) Interactive processes between mental and external operations in creative activity: a comparison of experts' and novices' performance. In *Proceedings of the 4th conference on Creativity & cognition table of contents,* pp 178-185, New York: ACM.

ISO 9241 (1997). Ergonomic Requirements for Office Work with Visual Display Terminals. ISO.

Jacobsen, N.E., Hertzum M. and John, B.E. (1998) The Evaluator Effect in Usability Tests. In *Proceedings of CHI'98.* New York: ACM Press.

Karat, C. M., Campbell, R., and Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In *Proceedings of CHI'92* (397-404). New York: ACM Press.

Kellogg, W. A. and Breen, T. J. (1987). Evaluating User and System Models: Applying Scaling Techniques to Problems in Human-Computer Interaction. In *Proceedings of CHI 1987* (303-308). New York: ACM Press.

Molich, R. (2000). *Usable Web Design* (In Danish). Ingeniøren|bøger.

NASA. Task Load Index, http://iac.dtic.mil/hsiac/Products.htm#TLX.

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

Nielsen, J. (1993). *Usability Engineering.* San Diego: Morgan Kaufmann.

Preece, J., Rogers, Y. and Sharp H. (2002). *Interaction Design: Beyond Human-Computer Interaction.* New York: John Wiley & Sons, Inc.

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

Rubin, J. (1994). *Handbook of Usability Testing: How to plan, design and conduct effective tests.* New York: John Wiley & Sons, Inc.

Urokohara, H., Tanaka, K., Furuta, K., Honda, M. and Kurosu M. (2000). NEM: "Novice Expert ratio Method" A Usability Evaluation Method to Generate a New Performance Measure. In *Extended Abstracts of CHI 2000,* pp 185-186, New York: ACM

8. ACKNOWLEDGEMENTS

We are grateful to the nurses as well as the personnel dealing with deployment and training for their participation in the empirical study. We also thank Kasper Hornbæk and Aage Nielsen for valuable comments concerning data collection and analysis.