Evaluating IT Systems for the Healthcare Domain: Longitudinal Usability Studies and Rapid Analysis Techniques

Jesper Kjeldskov, Mikael B. Skov, Jan Stage
Aalborg University, Department of Computer Science
Fredrik Bajers Vej 7, DK-9000 Aalborg, Denmark
{jesper, dubois, jans}@cs.aau.dk

ABSTRACT

This paper describes two studies evaluating the usability of IT systems for the healthcare domain. First we describe a longitudinal study of an Electronic Patient Record system, secondly, we describe a new technique for rapidly analyzing usability data.

Author Keywords

Usability, novices versus experts, rapid analysis

ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology

INTRODUCTION

As a part of the Digital Northern Jutland project, we have conducted a series of usability studies of IT systems for the healthcare domain between 2002 and 2004 in collaboration with Sygehus Vendsyssel and Virtual Centre for Health Informatics at Aalborg University. This paper describes and outlines some of our findings from two of these studies; a longitudinal usability study of Electronic Patient Records and a comparison between a rapid analysis technique (Instant Data Analysis) and traditional video data analysis.

A LONGITUDINAL USABILITY STUDY

In our first study, we investigated into the way novice and expert users experience the usability of an Electronic Patient Record system (EPR). Electronic Patient Records allow nurses and medial doctors to access and register information such as state, diagnosis, treatment, and medication of patients on a computer rather than on paper. The basic design of the study was to conduct two usability evaluations of the same EPR system with the same users one year apart in time. The first evaluation took place in May 2002 when the system was being taken into use at the hospital. The second evaluation took place in August 2003 when the users had used the system in their daily work for more than a year.

The Evaluated System

The EPR system used in our study was IBM IPJ 2.3, which is being used at Sygehus Vendsyssel. For the purpose of the usability studies, a test version of IPJ 2.3 was installed in our usability laboratory and configured to match the system used at the hospital

Novice Users

The first usability evaluation involved eight trained nurses. All eight nurses were women, aged between 31 and 54 years, their experience as nurses varied between 2 and 31 years. Prior to the first evaluation they had received between 14 and 30 hours of training in the IPJ system. They characterized themselves as novices or beginners in the IPJ system and in IT in general.

Expert Users

The purpose of the second evaluation was to study the usability of the EPR system after one year of use. Seven of the eight nurses in the first study were able to participate in the second evaluation. A participant with the same characteristics replaced the eighth nurse. 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 expert users.

The Two Usability Evaluations

Preparations: Prior to the first evaluation, we visited the hospital and had a number of meetings and discussions with the two persons who trained the nurses in using the IPJ system and dealt with the practical deployment of it. The purpose of this was to understand the work in the hospital wards related to the patient record and get an overview of the system and its parts. Based on this we made a number of overall scenarios for the use of the system and generated realistic test data.

Tasks: The purpose of the usability evaluations was to inquire into the extent to which the IPJ system supports nurses in solving work tasks that are typical for the hospital. Based on our scenarios, we designed three tasks centered on

Fourth Danish HCI Research Symposium, November 16, 2004
Aalborg University, Denmark.

the core purpose of the system such as retrieving information about patients, registering information about treatments, making notes, and entering measurements. The draft tasks were evaluated by nurses responsible for the training program.

Settings: All test sessions were conducted in our usability laboratory using a standard PC with a 19” screen matching the setup at the hospital.

Procedure: The test sessions were based on the think-aloud protocol as described by Rubin [[5]] and Nielsen [[3]]. In both evaluations, the eight test sessions were conducted over two days. The order of the nurses was random. One of the authors of this article was test monitor throughout all sixteen test sessions.

Data Analysis: The data analysis was done in August 2004, one year after the second evaluation. The two authors who did not serve as test monitor analyzed all sixteen 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. For each problem in the list there was 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 [[2]] and Nielsen [[3]]. Each evaluator also made a severity assessment of the usability problems as cosmetic, serious or critical [[2]].

The individual problem lists from the two evaluators were merged into one common list of usability problems. This was done in a negotiation process where the problems were considered one at a time. The evaluators also produced a log file of between two and four pages for each of the sixteen test sessions with the exact times and descriptions of the steps that the user goes through in order to solve each task. The log file also described whether the user solves each task, and to what extent the test monitor provides assistance.

RESULTS

Table 1 summarizes key results of problem identification for the novices and experts. 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 and a contingency analysis shows that this difference is significant (χ^2=8.489, *df*=1, *p*=0.0036).

Attributing severity to the identified usability problems, the highest experienced severity for each problem is used. We found that the novice subjects experienced significantly more serious problems than the experts (χ^2=4.296, *df*=1, *p*=0.0382), however no significant differences were found for the critical or cosmetic problems.

	Novice (N=8)	Expert (N=8)	Total (N=16)	χ^2	*p*
Critical	25 (21)	19 (17)	27 (21)	3.068 (2.487)	0.0798 (0.1148)
Serious	45 (29)	34 (23)	56 (32)	4.296 (2.564)	0.0382 (0.1093)
Cosmetic	13 (6)	10 (5)	20 (8)	0.409 (0.291)	0.5224 (0.5896)
All	83 (56)	63 (45)	103 (61)	8.489 (5.752)	0.0036 (0.0165)

Table 1. Total numbers of identified usability problems for the novices and experts. Numbers in parentheses show non-unique problems; problems experienced by at least two subjects.

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 [[1]] and thus a high reliability of the merged list of problems.

We also sought to explore differences and similarities in the problems identified by the two sets of subjects. Figure 1 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 and these mainly concern functionality and services that were not applied in the novice sessions, e.g. work task lists for nurses.

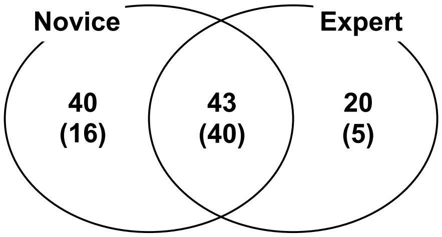

Figure 1. Distribution of the identified problems for the novices and experts. Numbers in parentheses show total numbers of problems subtracted unique problems.

Removing 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

INSTANT DATA ANALYSIS

Motivated by the challenges of analyzing usability evaluation data from our studies of Electronic Patient Record systems, we have developed a technique for reducing the efforts spent on analyzing data from usability evaluations: Instant Data Analysis (IDA). The aim of applying this technique is to make it possible to conduct an entire usability evaluation in one single day.

The IDA technique adopts the assumption that identifying the highest number of critical usability problems of a software product can lead to improved quality through redesign. The IDA technique is designed to be combined with the use of the well-established think-aloud protocol for user-based usability evaluations as described in for example [[4], [5]]. The technique can be applied to both laboratory-based and field-based think-aloud evaluations. The IDA technique exploits the fact that think-aloud usability evaluations typically involve a test monitor and a data logger with high level usability expertise. When conducting the evaluation, the test monitor and the data logger typically gain a strong insight into the evaluated system's key usability problems very quickly. While some of these problems may be captured by taking notes, much of this insight is often lost and needs to be reconstructed during later video data analysis. Rather than loosing this valuable moment of insight, the IDA technique extends the think-aloud sessions with a joint data analysis session.

Procedure

The use of the IDA technique follows immediately after think-aloud usability test sessions. Aiming at conducting the entire evaluation in one single day, 4 to 6 think-aloud sessions should provide a proper foundation for the analysis. During the usability test sessions, the data logger records incidents or problems. This will be used for the later problem identification and categorization. After the think-aloud sessions, the test monitor and the data logger conduct a one hour brainstorming and analysis session. The purpose of this session is to produce a list of usability problems as experienced by the 4 to 6 test subjects.

The roles of the test monitor and data logger during the data analysis are to articulate and discuss the most critical usability problems that they have identified during the think-aloud sessions. Also, they should rate the severity of each problem stating if it is, for example, critical, serious or cosmetic [[2]]. Assisting the brainstorming and analysis process, the test monitor and data logger may use printed screenshots of the system and notes taken by the data logger during the think-aloud sessions. The aim of the process is not to identify as many usability problems as possible, but to identify the most critical ones.

The analysis session is assisted by a facilitator. The role of the facilitator is to manage the brainstorming and analysis session, asking questions for clarification and writing all identified usability problems on a whiteboard/flip-over as they are presented by the test monitor and data logger. The facilitator should also make sure to keep an overview of the identified problems as the session progresses, categorizing them in themes, avoiding redundancy etc.

After the one hour brainstorm and analysis, the facilitator spends 1-1½ hour on his own writing up the contents of the whiteboard/flip-over into a ranked list of usability problems with short descriptions and clear references the system. Finally, the test monitor, data logger and facilitator run through the problem list together to ensure consensus.

COMPARING IDA TO VIDEO DATA ANALYSIS

We have evaluated the use of the proposed technique for Instant Data Analysis through a usability evaluation of a resource booking system for the healthcare domain.

Participants: The study included five test subjects of between 25 and 64 years of age. They were all staff at Sygehus Vendsyssel with practical experience ranging from 1 year to 37 years. The test subjects had all received training in the booking system. In addition, four trained usability researchers participated in different roles on evaluating the use of the IDA technique. All evaluators had significant previous experience usability evaluations. One researcher acted as test monitor during the test sessions with the five test subjects. A second researcher acted as data logger during the sessions writing down as much as possible during the tests. A third researcher observed the sessions and also logged data for supporting a later video analysis. Finally, a researcher observed the sessions and acted as facilitator in the IDA session.

Settings: The usability evaluation was conducted at the usability laboratory at Aalborg University. From the control room, the data logger could survey the subject room through one-way mirrors and by means of the motorized cameras. During the evaluation, the data logger took notes and created a preliminary log file. From the observation room, two researchers could observe the evaluation through one-way mirrors and on monitors relaying the screen image from the test PC and the cameras.

Procedure: The evaluation was conducted in one day (five hours). The individual sessions were structured by three tasks assignments given to the test subjects one at a time by the test monitor. During the evaluation, the test-subjects were thinking-aloud, explaining their interaction with the system and articulating their comprehension of the design.

Data analysis: The data from the usability evaluation sessions was analyzed independently by two teams of researchers applying a traditional video data analysis technique and the Instant Data Analysis technique respectively.

The Instant Data Analysis (figure 2) produced a list of usability problems ranked as critical, severe or cosmetic with approximately 2 lines of explanation. The total time spent using the traditional Instant Data Analysis technique amounted to 4 man-hours

Figure 2. Instant Data Analysis.

The analysis of the video data followed a standard approach to identifying usability problems. First, the preliminary log-files for each of the five test subjects created during the evaluation sessions were completed by looking through all videos. Following this, the video tapes were then examined thoroughly for identification of usability problems assisted by the log file and each usability problem was described in detail and ranked in relation to its severity.

The Video Data Analysis produced a detailed log file of the five evaluation sessions and a list of usability problems ranked as critical, severe or cosmetic with approximately 5-7 lines of explanation. The total time spent using the traditional Video Data Analysis technique amounted to approximately 40 man-hours.

Following the Instant Data Analysis and the video data analysis, the two lists of usability problems were merged in a collaborative effort. As a part of this, small variations in severity ratings were discussed until consensus had been reached.

FINDINGS

Comparing the IDA results with the results of the video data analysis approach, we found that the latter identified a total of 46 different usability problems where 12 were critical, 15 were serious, and 19 were cosmetic. In total, the two techniques identified a list of 62 different usability problems where 13 were critical problems, 22 were serious problems, and 27 were cosmetic problems.

Considering the identified problems, we found that both approaches assisted in identifying nearly all critical problems, where IDA identified 11 of the 13 (85%) critical usability problems whereas video data analysis identified 12 of the 13 (92%) critical usability problems.

The serious and cosmetic usability problems exhibited a different distribution between the two analysis techniques. Where the IDA technique identified 15 serious problems, a total of 22 serious problems were identified by the two approaches together. Thus, the IDA approach identified 68% of the serious problems found in total. On the other hand, the video data analysis also identified 15 serious problems (68%) meaning that eight serious problems were identified by both approaches. Considering the cosmetic problems, we found that the IDA technique identified 15 of the total 27 problems (56%). The 12 remaining cosmetic problems unidentified by IDA related primarily to specific interaction problems for the subjects typically only experienced by one of the five subjects. A total of 7 out of the 27 cosmetic problems (26%) were identified by both analysis approaches.

A high number of the usability problems identified in the video data analysis approach were experienced by only one subject test subject (26 problems of the total 46). It can be discussed whether these are really problems at all, or if they are noise added to the picture by non-generalizable subjective experiences of interaction with the system. Information about how many test subjects experienced the different usability problems was not included in the problem list generated from the IDA technique. But some of these 26 problems were also identified by the IDA approach. However, the majority of problems experienced only by one single test subject (16 of the 26) were *only* identified in the video data analysis and not in the instant data analysis. Thus, the use of the IDA approach allowed for the omission of a significant part of this noise.

CONCLUSIONS

This paper has reported from two usability studies of IT in the healthcare domain: a longitudinal usability study comparing the usability of an interactive system as experienced by novices and experts and a comparative study of two techniques for analyzing usability data.

In the first study, we observed that novices experienced more usability problems than the expert users of an Electronic Patient Record system. Yet a remarkably high number of problems were experienced both by novices and expert users. These problems were significantly more severe for the novices.

In the second study, we observed that using only 10% of the time required to do video data analysis, Instant Data Analysis helped identify 85% of the critical usability problems in the system being evaluated. At the same time, the noise of unique usability problems was significantly reduced.

REFERENCES

[1] Jacobsen, N.E., Hertzum M. and John, B.E. The Evaluator Effect in Usability Tests. Proc. CHI'98, ACM Press (1998).

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

[3] Nielsen, J. *Usability Engineering.* Morgan Kaufmann, San Diego (1993).

[4] Preece J., Y. Rogers H., Sharp D., Benyon D., Holland S. and Carey T. *Human-Computer Interaction.* Workingham, Addison-Wesley, 1994.

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