

Evaluating the Usability of Mobile Systems: Exploring Different Laboratory Approaches

Jesper Kjeldskov

Mikael B. Skov

Department of Information Systems
University of Melbourne, Australia
jesperk@staff.dis.unimelb.edu.au

Department of Computer Science
Aalborg University, Denmark
dubois@cs.auc.dk

Abstract

This paper addresses mobile system usability evaluations. Three different think-aloud evaluations of the same mobile system were conducted for the purpose of comparing the results by each approach. The evaluations spanned the use of test subjects with and without domain specific knowledge and the use of simple laboratory setups as well as high-fidelity simulations of the use context. The results show some significant differences between the results produced by the three approaches. However, it is indicated that while the recreation of a highly realistic use context resulted in a number of unique usability problems being identified, a large percentage of the usability problems found in total were identified already in the simple laboratory setups.

1 Introduction

Evaluating the usability of mobile systems constitute a potential challenge since their use is typically closely related to activities in their physical surroundings and often requires a high level of domain-specific knowledge (Nielsen, 1998). This can be difficult to recreate in a usability laboratory. Thus moving the evaluation into the real world may seem like an appealing approach. However, conducting usability studies in the field is also problematic. Access to real users and realistic settings can be difficult, data collection is complicated and means of control are limited.

This study is motivated by the need for evaluating the usability of a highly specialized mobile collaborative system supporting the coordination of safety critical work tasks on large container vessels (Kjeldskov and Stage, 2002). Evaluating this application was a challenge for a number of reasons. First of all, real users from the container vessels were not available for usability evaluations. Secondly, evaluating the system in the real world was not possible due to safety issues. Thus, the evaluation had to be done without going into the field and with limited access to prospective users. This challenged our ability to create a realistic laboratory setting.

2 Method

Three different evaluations of the mobile prototype were conducted for the purpose of comparing different approaches to creating realistic laboratory settings for mobile system evaluations.

2.1 Standard Laboratory with Non-Domain Subjects

Our first evaluation was conducted in a standard usability laboratory facilitating the observation of two physically separated subject rooms from a central control room through one-way mirrors.



Figure 1: Usability laboratory setup



Figure 2: Video from laboratory evaluation

Three two-subject teams, all computer science students, were given the task of coordinating a work task on board a fictive container vessel, communicating exclusively by means of textual commands on their mobile devices. The test subjects received a 15-minute introduction to the use context of the prototype application: the overall operation supported by the system, the basic concepts and maritime notions involved, the distribution of work tasks and present procedures of communication and coordination. Afterwards, one test subject was asked to act as captain on the bridge in one subject room while the other acted as officer on the fore mooring deck in the other room. The test subjects were asked to think-aloud during the evaluation. An evaluator located in each subject room observed the test subjects and asked them about their actions.

The laboratory setup consisted of two Compaq iPAQs connected through a wireless network displaying the interfaces for the officer on the fore mooring deck and the captain on the bridge respectively. Two A4 handouts depicted standard patterns of mooring and explained 10 basic concepts of the maritime context for quick reference. The test subjects were seated at a desk with the mobile device located in front of them (figure 1). Cameras mounted in the ceiling captured high quality video images of the evaluation sessions: overall views of the test subjects and close up views of the mobile devices. The video signals were merged into one composite signal and recorded digitally (figure 2).

2.2 Standard Laboratory with Domain Subjects

Our second evaluation applied the same laboratory setup, introductory procedure and tasks as described above. However, for the purpose of increasing the realism of the evaluation, we altered the experiment in two ways. First, we brought in prospective users from the nearby Skagen Maritime College. All test subjects were thus skilled sailors with practical experience with the operation of large vessels. Secondly, we introduced a simple paper mock-up of a ship in harbor and central instruments on the bridge. Apart from introducing a more realistic context of use, the purpose of this mockup was to supply the test subjects with a tool for explaining their strategies and actions. The test subjects acting as captains were thus asked to operate the controls of the mockup as they would operate the controls on the bridge in the real world and use the model of the ship to illustrate the process as it developed over time.

2.3 Advanced Laboratory with Domain Subjects

Our third evaluation took place in a temporal usability laboratory at the simulation division of Svendborg International Maritime Academy and used their state-of-the art ship simulator for

creating a highly realistic but yet safe and controllable experimental setup. The ship simulator consisted of two separate rooms: an operator room (also resembling the fore mooring deck) and a simulated bridge fully equipped with realistic controls and instruments (figure 3). The simulator was set up to imitate the operation of a large vessel in challenging weather and traffic conditions corresponding to a real world situation. The academy provided test subjects with practical experience on the operation of large commercial and military vessels. Three teams of two test subjects were given the introduction and overall task described above. During the evaluation, the test subject acting as captain had to consider all aspects of maneuvering the ship as well as communicating with personnel, harbor traffic control etc. and taking into consideration the movements of other vessels. Two evaluators located on the simulated bridge and operator room respectively observed the test subjects and asked questions for clarification. As in the standard laboratory studies the prototype setup consisted of two Compaq iPAQs and four high quality video images of the evaluation sessions were recorded digitally (figure 4).

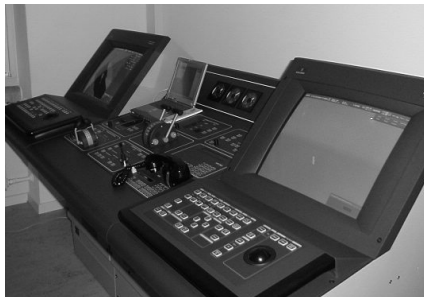


Figure 3: The high-fidelity ship simulator



Figure 4: Video from evaluation in simulator

2.4 Data Analysis

The data from the evaluation sessions consisted of three video recordings with the total of 18 test subjects. The analysis of this data aimed at creating three lists of usability problems experienced by the test subjects: one list for each of the three studies. The videotapes were analysed in three steps. First, problems experienced on deck were identified by examining the videos while listening to audio from one subject room only. Secondly, the same was done for identifying problems experienced on the bridge. Finally, the videotapes were examined listening to audio from both subject rooms simultaneously. This analysis was done in a collaborative effort between the two authors allowing an immediate discussion of each identified problem.

3 Results

Figure 5 outlines the distribution of the problems identified in the three evaluations. The three blocks signify the standard laboratory with non-domain subjects (#1), standard laboratory with domain subjects (#2), and advanced laboratory with domain subjects (#3) studies, each divided into problems identified on the bridge and on the deck respectively. Each column represents a unique usability problem and a black box means that the problem was identified by that approach.

Totally, 58 unique usability problems were identified in the three evaluations. 7 problems were identified in all three evaluations and on both the bridge and on the deck. Some of these problems were related to interaction issues e.g. finding out which elements on the screen to interact with. Another general problem found was that many test subjects did not see all relevant state changes

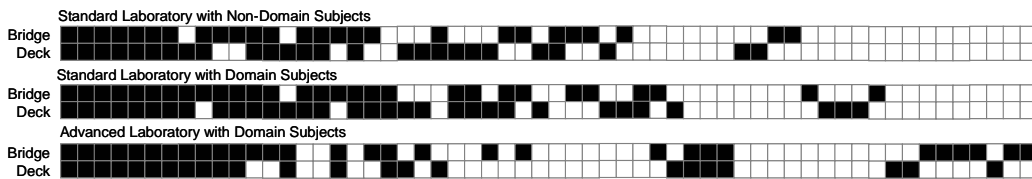


Figure 5: Distribution of identified usability problems in the three evaluations

in the system, which sometimes caused them to miss commands or confirmations. Other general problems were related to the correlation between the representation of the ship and ongoing activities in the system and real activities on the ship.

The study revealed no significant difference in the number of identified problems for the three evaluations. The standard laboratory with non-domain subjects in total identified 37 usability problems, where the standard laboratory with domain subjects identified 40, and the advanced laboratory with domain subjects identified 36. However, the number of unique problems identified was significantly higher for the advanced laboratory study, identifying 12 unique problems of which 3 were discovered both on the bridge and on the deck. In comparison, the standard laboratory with non-domain subjects identified 5 unique problems and the standard laboratory with domain subjects 6. On an overall level, it is noticeable that 35% of the total number of problems was identified by domain subjects only, stressing the importance of including prospective users. On the importance of a realistic use context, however, it is noticeable that 80% of the total number of problems was actually identified outside the simulator. On the other hand, almost 40% of the total number of problems was not encountered by domain users in the simulator. This is an interesting issue as it is difficult to say whether these problems are simply not relevant since they were not discovered in the most realistic setup or if the simulator setup infused so much complexity on the session that some shortcomings might have ignored.

4 Discussion

The key literature on usability testing and engineering typically states that usability evaluators should minimize their influence on the conduction of the test and on the data analysis, e.g. (Nielsen, 1993; Rubin, 1994). This usually means assigning real users to the test in order to identify realistic and relevant problems, and carefully elaborating assignments in order to focus on relevant aspects of the system. However, several studies have shown that various aspects of an evaluation influence the results, e.g. the evaluator-effect (Jacobson, Hertzum, and John, 1998), the number of test subjects (Lewis, 1994), and the level of investigator intervention (Held and Biers, 1992). In our study, we have deliberately explored the influence of changing different settings of the evaluation: varying the test subjects by including non-domain users and applying different levels of contextual realism to the laboratory setup.

While no significant difference was found regarding the number of problems identified in the three studies, analyzing our results qualitatively, a number of interesting differences emerge. First of all, the character of the unique problems identified in each study was different. While the 5 unique problems identified by non-domain users could generally be related to lack of knowledge about the use context, many of the 18 unique problems identified by domain subjects in the standard and advanced laboratory studies combined were concerned with highly relevant issues such as the representation of the task in the system and lack of flexibility in coordination and communication. E.g. some of the domain subjects wanted to specify commands more detailed than supported by

the system. Some of the 12 unique problems identified in the advanced laboratory were furthermore related to critical issues such as not being able to cancel commands when changed conditions in the simulation required this. This turned out to be a critical problem since the captain had to apply different means of communication in order to achieve his goal, which resulted in further problems regarding the representation of the real world in the system. No such situations occurred in the standard laboratory and none of the non-domain subjects wanted to cancel commands or expressed that not being able to do so could be a problem. Secondly, the realism of the environment of the advanced laboratory implied that the test subjects had to operate other systems and consider other information apart from the evaluated mobile prototype. Hence, the test subjects' attention towards the mobile device in the advanced laboratory was lesser than in the standard laboratory studies. This resulted in test subjects often missing updates or changes on the display of their mobile device in advanced laboratory sessions while this was not a significant problem to users in the standard laboratory. Also the realism of the context in terms of weather and traffic conditions induced by the simulator made an impact on the results of the study causing some of the test subjects to apply approaches to the operation and request procedures, which were not supported by the system. None of the test subjects in the standard laboratory evaluations experienced that problem.

Our study indicates that central usability problems of a mobile system can be identified in laboratory settings. While it is shown that including prospective users and recreating realistic contexts support the identification of qualitatively different problems, it is also shown that a large number of a mobile system's usability problems can be found in simpler laboratory approaches. The results of our study are limited in a number of ways. First, the number of test subjects applied in each evaluation limits their general validity. In order to be able to make general recommendations, the study needs to be replicated with more subjects and probably varying the system. Secondly, the test subjects were not required to be as mobile in the evaluations as they would have been in a corresponding real world situation. Finally, we have not at this point assessed the severity of the identified problems. This means that we cannot conclude on the implications of the discovered problems and also limits our comparison of the three studies. This analysis is forthcoming.

References

- Held, J. E. and Biers, D. W. (1992) Software Usability Testing: Do Evaluator Intervention and Task Structure Make Any Difference? *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica, HFS, pp. 1215 – 1219
- Jacobson, N. E., Hertzum, M., and John, B. (1998) The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp. 1336 – 1340
- Kjeldskov, J. and Stage, J. (2002) Designing the User Interface of a Handheld Device for Communication in a High-Risk Environment. *Adjunct Proceedings of the 7th ERCIM Workshop on User Interfaces for All*, Paris, France
- Lewis, J. R. (1994) Sample Sizes for Usability Studies: Additional Considerations. *Human Factors*, 36(2), pp. 368 – 378
- Nielsen, C. (1998) Testing in the Field. Proceedings of the third Asia Pacific Computer Human Interaction Conference. Werner, B. (ed.), *IEEE Computer Society*, California, pp. 285-290
- Nielsen, J. (1993) *Usability Engineering*. Boston, Academic Press
- Rubin, J. (1994) *Handbook of Usability Testing – How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons