*Human-Computer Interaction -- INTERACT'03*
*M. Rauterberg et al. (Eds.)*
*Published by IOS Press, (c) IFIP, 2003, pp. 663-670*

# Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluations of a Mobile System

**Jesper Kjeldskov and Mikael B. Skov**

Department of Computer Science, Aalborg University,
Fredrik Bajers Vej 7, 9220 Aalborg East, Denmark

{jesper, dubois}@cs.auc.dk

**Abstract:** This paper addresses the issue of creating a realistic laboratory setting when evaluating the usability of mobile systems. Three laboratory-based think-aloud evaluations of the same mobile system were designed and conducted for the purpose of comparing the impact of different approaches to creating realistic laboratory environments on the results subsequently produced. The three evaluations spanned the use of test subjects with and without domain specific knowledge and the use of low and high fidelity simulations of the use context. The results show similarities and differences in the results of the three evaluations. First, an equal number of usability problems are identified in the evaluations, and a standard laboratory involving non-domain subjects can identify severe problems. On the other hand, employing domain test subjects in a simulated laboratory seems to generate some additional problems which would be difficult to identify in the standard laboratory.

**Keywords:** Evaluation, usability, mobile systems, safety-critical, think-aloud

## 1 Introduction

Usability evaluations have proven to be invaluable tools for assessing computerized systems according to user mistakes and satisfaction (Molich, 2000; Hodgson and Ruth, 1985). However, evaluating the usability of mobile systems constitutes a potential challenge since the use of such systems is typically closely related to activities in their physical surroundings and often requires a high level of domain-specific knowledge (Nielsen, 1998).

Conducting mobile device usability evaluations in the field raises a number of potential problems. First, the use contexts of mobile devices are often highly dynamic involving several temporally and physically distributed actors. Additionally, mobile systems for highly specialized use contexts such as e.g. safety-critical or hazardous environments may prohibit exploratory usability evaluations since errors involving risks for people and equipment cannot be tolerated. Finally, field evaluations complicate data collection and limit means of control since users are moving physically in an environment with a number of unknown variables potentially affecting the setup.

When conducting usability studies in laboratory-based settings, experimental control and collection of high quality data is typically not a problem. Fundamentally, laboratory usability evaluations may be either expert, theoretical, or user-based (Henderson et. al., 1995) or a combination of these (Cuomo, 1994). The similarities and differences between theoretical evaluations, e.g. heuristic inspection, and user-based evaluations, e.g. usability testing with think-aloud, have been shown for standard desktop applications, cf. (Karat et. al., 1992), and for web-sites, cf. (Kantner and Rosenbaum, 1997). It is generally acknowledged that user-based evaluations tend to find a higher number of problems and more relevant problems (Karat et. al., 1992). On the other hand, user-based evaluations tend to be more time consuming than theoretical evaluations, cf. (Nielsen and Molich, 1990). We studied user-based and theoretical evaluations for collaborative mobile systems and found that user-based evaluations seem to find more problems than theoretical evaluations (Kjeldskov and Skov, 2003).

Laboratory evaluations of mobile systems raise a number of challenges. First, the relation between the system and activities in the physical surroundings can be difficult to capture in expert evaluations such as heuristic evaluation or recreate realistically in a usability laboratory. Secondly, working with systems for highly specific domains (Kjeldskov and Skov, 2003; Luff and Heath, 1998), laboratory studies may be impeded by limited access to prospective users on which such studies rely. While benefiting from the advantages of a controlled experimental space, evaluating the usability of mobile systems without going into the field thus challenges established methods for usability evaluations in controlled environments: How can realistic laboratory settings be created that capture a mobile system use context? And how do different levels of realism applied to laboratory-based evaluations influence the results produced? How much realism is enough for identifying the usability problems of a mobile system?

We address the above questions by exploring ways of creating laboratory settings with different levels of realism for evaluating the usability of mobile systems. Section 2 outlines the background for the study including the case and the evaluated mobile system. Section 3 presents the method behind the study describing how the evaluations were set up and carried out. Section 4 highlights the key results of the study, and finally section 5 discusses the results and their potential implications and limitations.

# 2 Background

This study was motivated by the need for evaluating the usability of a mobile system for professional workers performing safety-critical collaborative work task on board very large container vessels. Evaluating this application was a challenge for a number of reasons. First, prospective users from the container vessel are not easily available for this kind of 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, consequently challenging our ability to create a realistic laboratory setting.

## 2.1 The Use Context

The mobile system to be evaluated was developed for supporting work activities on board large container vessels (their size equivalent of app. 3½ football fields). The operation of such vessels requires workers to be highly mobile and physically distributed. Typically, the crew number is not very high and hence people are assigned to various tasks on different locations on the ship depending on the situation: cruising at sea, departing from the quay etc. Work activities on large container vessels are typically safety-critical and involve high risks in the case of errors. Especially when maneuvering inside a harbor, erroneous actions may cause serious material damage and possible injuries to personnel or loss of human life. Thus, mobile systems for supporting these work activities have to be carefully evaluated.

We designed an experimental prototype to support the work task “letting go the lines” (Kjeldskov and Stage, 2002) based on ethnographic studies of work activities on a container vessel (Andersen and May, 2001; Nielsen, 2000). Here, the first step in leaving the quay is letting go the mooring lines holding the vessel in a fixed position. However, as physical space is restricted and means for precise maneuvering are limited, all lines cannot simply be released simultaneously. When a line is let go, it will remain in the water for a period of time during which no means of propulsion is available due to the risk of lines getting sucked in and wrapped around the propeller or thrusters. Instead the vessel can be pulled ahead or astern by means of the remaining lines.

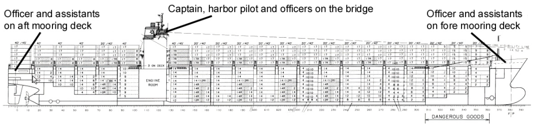

**Figure 1**. Sally Maersk - one of the world’s largest container vessels in operation

Due to the huge size of container vessels, the work tasks involved are distributed among a number of actors located at strategic positions (figure 1). These actors are all highly mobile throughout the whole operation. On the bridge, chief officers control the rudder, propeller and thrusters. Fore and aft, the first and second officers control the winches for heaving in the lines. Ashore, two teams of assistants lift the lines off the bollards. The challenge consists of bringing the vessel away from the quay sideways without running aground in the shallow water behind it or colliding with the ship at quay in front of it. Because of wind, current, temporal lack of propulsion and poor visual view from the bridge, this operation is not trivial and relies on ongoing communication and careful coordination.

At present this coordination is primarily based on oral communication following well established formalized procedures. While people on the bridge can see and hear each other, personnel on deck are out of direct visual and audio contact and have to communicate with the captain via walkie-talkies. In order to carry out the operation of departure, the captain needs an overview and total control over the propulsion, direction and mooring of the ship. While information about the rudder, propeller and thrusters are available on dedicated instruments no information about mooring is facilitated. At present this only exists as a mental model in the head of the captain based on his perception of the ongoing communication between bridge and deck. As this mental model is highly sensitive to errors or misunderstandings in the communication, and since disparity between the captain's mental model and the real world may cause wrong decisions, considerable cognitive resources are spend on establishing and maintaining common ground among the cooperating actors (Clark and Schaefer, 1989).

## 2.2 Present Means for Coordination

Although well-established plans and formalized procedures for communicative coordination exist, increasing size of vessels and use of technology add to the complexity of the work tasks to be coordinated. Thus a strong motivation exists for supporting communication and coordination of work tasks in this domain better. Specifically, the field studies revealed that sound quality is often poor and that communication is not persistent.

Asynchronous text-based messaging is a flexible, ubiquitous and persistent communication channel requiring low cognitive overhead, e.g. (Churchill and Bly, 1999). So it was the thesis of the development team, that a text-based communication channel on mobile devices could eliminate or reduce some of these limitations.

## 2.3 The Prototype Application

Based on the analysis described above, a prototype of the "Handheld Maritime Communicator" was designed and implemented (Kjeldskov and Stage 2002). The prototype setup consisted of three iPAQ 3630 connected through an IEEE 802.11b 11Mbit wireless TCP/IP network. One device was indented for the captain on the bridge while the other two were indented for 1st and 2nd officers on the fore and aft deck respectively. The prototype was implemented using Microsoft Embedded Visual Basic.

This mobile system gives the distributed actors on the container vessel access to a mobile text-based communication channel and provides a graphical representation of the ship and its mooring lines. The interface is divided into four sections: Pictogram of the ship and lines, list of completed communication threads, list of ongoing threads of communication, and list of unexecuted commands.

Get ready fore and aft — Ready fore
Ready aft
Single up 1 and 1 fore — 1 and 1 fore
Single up 1 and 1 aft — 1 and 1 aft

Let go bow line [ok] Line let go [ok]
Heaving in.. 0:49
Let go stern spring [ok] Spring let go [ok]
Heaving in.. 0:03
Let go stern line 0:01

Let go bow spring
Let go all fore

**Figure 2**. Compaq iPAQ handheld computer and a screen shot from the Handheld Maritime Communicator

At the bottom of the screen, unexecuted commands and confirmations are displayed on a list. The order of the list corresponds to the standard sequence of the operation and commands only appear when appropriate. By default, the most likely next step of the operation is highlighted. Commands can be browsed and executed (send) with the five-way key on the device.

# 3 Method

Three different studies of the mobile prototype were conducted for the purpose of identifying usability problems to be addressed in the next iteration of design and at the same time evaluating and comparing different approaches to creating realistic laboratory settings for mobile system evaluations: Standard lab w/ non-domain subjects, standard lab w/ domain subjects, and advanced lab w/ Domain Subjects.

## 3.1 Standard Lab w NonDomain Subjects

We focused on evaluating the usability of the mobile prototype with a minimal effort spent on creating a realistic setting. As in traditional laboratory-based think-aloud usability evaluations, context was primarily provided by the tasks. However, realism was also supported through physical separation of test subjects and an introduction to the use context and a possible use scenario.

The first study was conducted in a standard usability laboratory consisting of two subject rooms and a control room. From the control room, both subject rooms could be surveyed through one-way mirrors and by means of remotely controlled motorized cameras mounted in the ceiling. From each subject room, it was not possible to look into the other. For our specific experiment, subject room #1 resembled the bridge while subject room #2 resembled the fore deck.

Three teams of two test subjects, all CS students, were given the task of letting go the lines before departure of a large vessel. The test subjects received a 15-minute joint oral introduction to the use context of the prototype application and were presented with a use scenario. This was supported by a number of illustrations on a whiteboard (figure 3). The introduction and use scenario covered the whole operation to be supported, the basic concepts and maritime notions involved, the distribution of work tasks and present procedures of communication and coordination (as described in section 2). Following this, one person was asked to act as the captain on the bridge while the other acted as officer on the fore mooring deck.

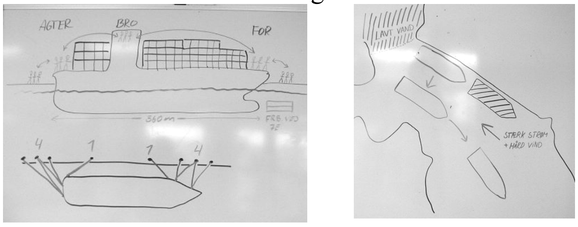

**Figure 3**. Introduction to use context and a possible scenario drawn on whiteboard

During the evaluation, the test subjects were asked to think-aloud, explaining their comprehension of and interaction with the prototype. An evaluator located in each test room observed the test subjects and frequently asked them about their actions. On a video monitor facing away from the test subjects, the evaluators could see a close up view of the mobile device as well as the activities in the other subject room for sake of overview.

The laboratory setup consisted of two Compaq iPAQs and a PocketPC emulator on a laptop PC connected through a wireless network. The iPAQs displayed the interfaces for the officer on the fore mooring deck and the captain on the bridge respectively. The laptop displayed the interface for the officer on the aft mooring deck and was operated by one of the evaluators. Two A4 handouts depicted standard patterns of mooring and explained 10 basic concepts and notions of the maritime context for quick reference if necessary.

The test subjects were seated at a desk with the mobile device located in front of them. Video cameras mounted in the ceiling captured high quality video images of the evaluation sessions. Two cameras captured global views of the test subjects and two cameras captured close up views of the mobile devices. In order to ensure good video images of the displays, the test subjects were asked to keep the mobile devices within a delimited area, drawn on a white piece of paper taped to the desk. The four video signals were merged into one composite signal and recorded digitally (figure 4). Audio from the two subject rooms was recorded on separate tracks for later mixing and potential separation during playback.

**Figure 4**. Video from evaluation in usability lab

## 3.2 Standard Lab w/ Domain Subjects

In our second think-aloud evaluation, we used the same laboratory setup and introductory procedure as described above. However, effort was put into increasing the realism of the setting by altering the experiment in two ways. First, prospective users from Skagen Maritime College were brought in. All test subjects were educated and practically skilled sailors experienced with the operation of large vessels including hands-on experience with the task of letting go the lines. Secondly, we introduced a simple mock-up of the ship (figure 5).

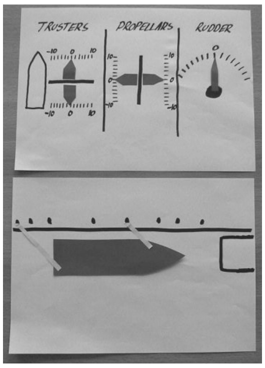

**Figure 5**. Mock-up of the bridge, ship and mooring

The mock-up consisted of central instruments on the bridge for controlling the thrusters, propellers and the rudder as well as a model of the ship and mooring lines placed on a schematic drawing of the harbor. 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 on the bridge were thus asked to operate the controls of the mockup as they would operate the controls on the bridge in the real world and to use the model of the ship to illustrate the overall process of departing from the harbor as it developed over time.

As in the first usability laboratory study, three teams of two test subjects were given the task of letting go the lines before departure of a large vessel. During the evaluation, the test subjects were asked to think-aloud, under the observation of an evaluator asking questions for clarification.

## 3.3 Advanced Lab w/ Domain Subjects

We aimed at evaluating the prototype in the hands of prospective users in a highly realistic laboratory setting. Accomplishing this, we established a temporary usability laboratory at the simulation division of Svendborg International Maritime Academy and used their state-of-the art ship simulator for creating a realistic but yet safe and controllable experimental setup. Again, we brought in educated and practically experienced prospective users as test subjects, this time from the academy running the simulator facility.

The ship simulator consisted of two separate rooms: a bridge and a control room. The bridge was fully equipped with controls for thrusters, propellers, rudder etc. as well as instruments such as Doppler log, echo sounder, electronic maps, radars, VHF radio etc. From the control room, simulator operators could see the bridge on a closed circuit video surveillance system. The computer application driving the simulation facilitated a high-fidelity interactive scenario of the operation of any computer-modeled vessel at any modeled physical location. Also weather and dynamic traffic conditions could be included into the scenario.

**Figure 6**. The high-fidelity ship simulator

The simulator was set up to imitate the operation of a large vessel in challenging weather and traffic conditions in Felixstowe harbor corresponding to a real world situation observed during our field studies (Nielsen, 2000).

Three teams of two test subjects were given the task of letting go the lines and departing from harbor using the presented prototype for communication between bridge and deck. One test subject acted as captain on the simulated bridge while the other acted as 1st officer on the fore deck in the neighboring simulator control room. For simplicity, commands targeted at the 2nd officer on the aft deck were fed directly into the simulation and feedback was given by the simulation operator.

Carrying out the operation, the test subject acting as captain had to consider all aspects of maneuvering the ship. This included controlling the rudder, propellers and thrusters as well as communicating with personnel on the ship, harbor traffic control, etc. and taking into consideration the movements of other vessels. The primary task of the 1st officer on deck was to orally forward commands executed by the captain via the mobile device prototype to the operator of the simulation (impersonating as the team of assistants carrying out the actual tasks) and report back to the captain. The operator would then enter the commands into the simulation, and report to the 1st officer when the requested operations (such as letting go a line) had been carried out.

**Figure 7**. Test setup in the simulator

During the evaluation, the test subjects were asked to think-aloud, explaining their comprehension of and interaction with the prototype. Two evaluators located on bridge and deck respectively observed the test subjects and asked questions for clarification. On a video monitor facing away from the test subject, the evaluator on the deck could see a close up view of the test subject's mobile devices as well as an overview of the bridge.

As in the standard laboratory studies, the prototype setup consisted of two Compaq iPAQs and a PocketPC emulator on a laptop PC connected through a wireless network. High quality video images were captured of the evaluation sessions. An already installed stationary surveillance camera captured an overall view of the simulated bridge while close-up views of the test subject's interaction with the prototype and other controls on the bridge was captured by the evaluator using a hand-held

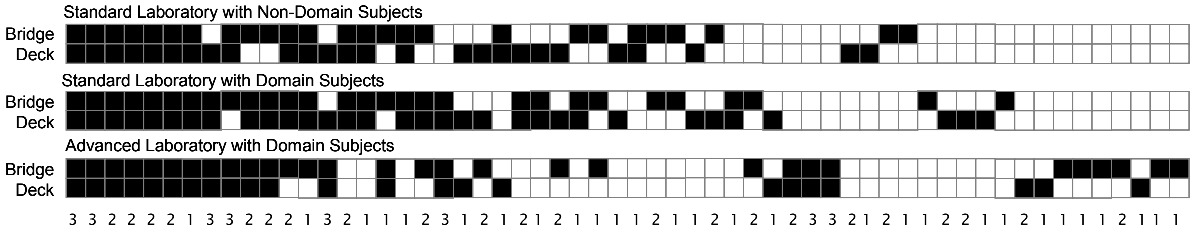

**Figure 8**: Distribution of the 58 identified usability problems in the three evaluations; a black box means that the usability problem was identified by the corresponding evaluation. The three horizontal blocks outline the problems identified in each of the three settings where each block is divided into problems identified on the bridge and on the deck respectively. The problem list is sorted by the number of configuration identifying the problem and severity of the problem. The bottom line indicates the severity of the identified problem (1=cosmetic, 2=serious, and 3=critical).

camera. In the room resembling the fore mooring deck, a camera was set up on a stand for capturing an overall view of test subject as well as the operators of the simulator. As in the usability laboratory, the test subject acting as officer on deck was seated at a desk with the mobile device located in front of him. Again, the device had to be kept within a delimited area drawn on a white piece of paper taped to the desk on order to ensure good video images of the display. The four video signals were merged into one composite signal and recorded digitally and audio from the two rooms was recorded on separate audio tracks.

### 3.4 Data Analysis

The data analysis aimed at creating three lists of usability problems identified in each of the three approaches classified as cosmetic, serious, or critical (Molich, 2000). The usability testing consisted of three sessions of video recordings depicturing the test sessions with the 18 test subjects.

All videotapes were afterwards examined and analyzed by the two authors of this paper in three steps. First, problems for the test subject on the fore deck were identified by going through the video while only listening to speak on the fore deck. Secondly, problems for the test subject on the bridge were identified by going through the video while only listening to speak on the bridge. Thirdly, the videotapes were lastly examined listening to speak on the fore deck and the bridge simultaneously in order to identify further usability problems.

The video examination and analysis was done in a collaborative effort between the two authors allowing immediate discussions of identified problems. The examination resulted in a list of usability problems.

## 4 Results

The three evaluations resulted in identification of 58 unique usability problems. The distribution of the identified problems is outlined in figure 8.

The standard laboratory with non-domain subjects identified 37 (63.8%) of the 58 problems and 14 (24.1%) of these problems were experienced on both the bridge and on the deck. The standard laboratory with domain subjects identified 40 (69.0%) of the 58 problems and 21 (36.2%) of these problems were experienced on both the bridge and on the deck. The advanced laboratory with domain subjects identified 36 (62.1%) of the 58 problems and 17 (29.3%) of these problems were experienced on both the bridge and on the deck. Seven (12.1%) problems were identified in all three evaluations on both the bridge and on the deck. Some of these problems relate to interaction issues, e.g. nearly all test subjects had problems about which elements to interact with on the screen whereas a few relate to the correlation between the representation of the ship on the system and real activities on the ship, e.g. many test subjects could not state the status of commands they had issued.

| | #1 | #2 | #3 |
|---|---|---|---|
| No. of problems | 37 (64%) | 40 (69%) | 36 (62%) |
| - Critical | 5 (63%) | 6 (75%) | 8 (100%) |
| - Serious | 16 (73%) | 17 (77%) | 14 (64%) |
| - Cosmetic | 16 (57%) | 17 (61%) | 14 (50%) |
| Unique problems | 5 (24%) | 5 (24%) | 11 (52%) |
| - bridge and deck | 1 (20%) | 1 (20%) | 3 (60%) |

**Figure 9**: Characteristics of identified problems in the three evaluations

Looking at the severity of the identified problems we find that eight (13.8%) of the 58 problems were critical problems, 22 (37.9%) of the 58 problems were serious problems, and 28 (48.3%) of the 58 problems were cosmetic problems. Figure 9 outlines characteristics of the identified problems in all three evaluations.

The distribution of problems on severity furthermore reveals that all three evaluations identify a large portion of the critical and serious problems. However, the advanced laboratory is able to reveal all eight critical problems. The number of unique identified problems for each of the evaluations is higher for the advanced laboratory where 11 of problems are unique to that approach and 3 of these were discovered both on the bridge and on the deck.

Five usability problems were identified by the non-domain subjects only and thus unique to this evaluation. These problems primarily relate to the conduct of the assignment, e.g. more of these test subjects did not know the correct or exact order in which commands should be executed. These kinds of issues caused no problems to the domain subjects. It should also be noticed that 20 of the identified usability problems were only detected by domain subjects in either the standard or advanced laboratory. In summary, almost 35% of the problems were not identified by non-domain subjects.

11 usability problems were unique to the advanced laboratory with domain subjects which constitute almost 1/3 of the total identified problems for that approach. These problems concern the representation of the task in the system and lack of flexibility, e.g. more of the domain subjects wanted to specify in more details how they wanted to depart the harbor. However, this was not possible in the system. Furthermore, some of these problems relate to the lack of being able to cancel actions, e.g. one test subject lost complete overview of what was going on since he had to cancel one action.

# 5 Discussion

Usability engineering literature states that usability evaluators should strive to minimize their influence on the usability test conduction (Nielsen, 1993). One should also strive to avoid other actions that would influence the result of the usability evaluation, e.g. implicit or explicit guidance of the test subject during the evaluation (Molich, 2000). However, several studies have shown that various aspects of the test influence the results, e.g. number of test subjects (Lewis, 1994), level of investigator intervention (Held and Biers, 1992). We deliberately try to explore the influence of changing (and affecting) different settings of the evaluation.

Previous studies relate to and discuss different usability evaluation approaches and methods. Karat found that user-based evaluations generate significantly more (and more relevant) usability problems that theoretical evaluations (Karat et. al., 1992). Henderson made an analogous examination that compares four different user-based evaluation methods and found that the usability testing with think-aloud generates most usability problems (Henderson et. al., 1995). Both studies, however, conclude that usability testing with think-aloud is the most expensive in terms of time consumption.

We cannot identify any major differences in the number of identified usability problems in the three evaluations. In this sense, it seems that inexperienced non-domain test subjects find as many usability problems as domain test subjects. The environment for the evaluation seems to have little or no impact on the number of identified usability problems where we identified 37 and 40 problems in the standard laboratory evaluations and 36 problems in advanced laboratory. Our study also seems to indicate that the advanced laboratory facilitates the identification of a higher number of unique problems.

The distribution of the usability problems is somewhat different for the three evaluations as illustrated in figures 8 and 9. It seems that we were able to identify some severe usability problems with the non-domain subjects in the standard laboratory, but we were not able to identify all critical problems. The two critical problems discovered by only the advanced laboratory evaluation were caused by interference by the environment during the test. Secondly, all three evaluations result in the identification of unique problems for that particular evaluation. E.g. the non-domain test subjects experienced problems in understanding the correct order of commands. This was no problem to the domain subjects.

Looking at the unique problems in the standard and advanced laboratories with domain subjects, we identify a number of interesting issues. First, more of the domain subjects needed to cancel already issued commands. This was not possible in the tested system. This turned out to be a critical problem in the evaluation in the advanced laboratory since the captain in the bridge had to apply different means of communication in order to cancel the command. None of the non-domain subjects wanted to cancel commands or issued that this could be a problem. Secondly, the realism of the environment of the advanced laboratory resulted in the test subjects (especially on the bridge) having to operate and consider other information resource, e.g. the captain had to operate and navigate the actual controls of the vessel. Hence, the users' attention on the system was often disturbed which resulted in the test subjects often missing updates or state changes in the investigated system. Thirdly, the complexity of the task imposed by the advanced laboratory, e.g. the conditions of the harbor in terms of other ships etc, made an impact on the results of the study. Some of the test subjects on the bridge wanted to apply a different approach to the command of letting go the lines. The system was restricted in that operation where certain sequential procedures could not be avoided. However, the conditions imposed by the advanced laboratories made the test subjects request different procedures in the operation. None of the test subjects in the standard laboratory evaluations discovered that problem.

The general validity of the results of our study is limited in a number ways. First, the number of test subjects applied in each evaluation implies that we can only explore the qualitative issues of changing the setting for usability evaluations. Our study can hopefully set out avenues for further research. Secondly, our introduction to the test subjects constituted an additional influence source on the results since this introduction was significantly important to the non-domain test subjects as they had no prior understanding of the domain. We strived to standardize this introduction as illustrated in section 3. Thirdly, the mobility aspect of our system was not directly evaluated in our study as the test subjects were situated in the same place during the evaluation. Brewster has addressed this aspect and found that subjects have more interaction problem when walking with a mobile device than when sitting down (Brewster, 2002).

# References

Andersen, P. B. and May, M. (2001) Tearing up Interfaces. In Liu K. et al (eds.) Information, organization and technology. *Studies in organizational semiotics*. Boston, Kluwer

Brewster, S. (2002) Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, vol. 6, pp. 188 - 205

Churchill E. F. and Bly S. (1999) It's all in the words: Supporting work activities with lightweight tools. In *Proceedings of ACM Siggroup'99*, pp. 40 - 49

Clark, H.H. and Schaefer, E. F. (1989) Contributing to discourse. *Cognitive Science*, vol. 13, pp. 259 - 294

Cuomo, D.L. (1994) A Method for Assessing the Usability of Graphic, Direct-manipulation Style Interfaces. *International Journal of Human-Computer Interaction*, vol. 6, pp. 275-297

Held, J. E. and Biers, D. W. (1992) Software Usability Testing: Do Evaluator Intervention and Task Structure Make Any Difference? In *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica: HFS, pp. 1215 – 1219

Henderson, R., Podd, J., Smith, M., and Varela-Alvarez, H. (1995) An Examination of Four User-Based Software Evaluation Methods. *Interacting with Computers*, vol. 7, no. 4, pp. 412-432

Hodgson, G.M. and Ruth, S.R. (1985) The Use of Menus in the Design of On-Line Systems: A Retrospective View. *SIGCHI Bulletin*, vol. 17, pp. 16-22

Kantner L. and Rosenbaum, S. (1997) Usability studies of WWW sites: Heuristic Evaluation vs. Laboratory Testing. In *Proceedings of the 15th annual international conference on Computer documentation*, pp. 153 – 160

Karat, C. M., Campbell, R., and Fiegel, T. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In *Proceedings of the ACM CHI92 Conference*. Reading, MA: Addison-Wesley, pp. 397 – 404

Kjeldskov, J. and Skov, M. B. (2003) Evaluating the Usability of a Mobile Collaborative System: Exploring Two Different Laboratory Approaches. In *Proceedings of the 4th International Symposium on Collaborative Technologies and Systems*, pp. 134 - 141

Kjeldskov, J. and Stage, J. (2002) Designing the User Interface of a Handheld Device for Communication in a High-Risk Environment. In *Proceedings of the 7th ERCIM Workshop on User Interfaces for All*

Lewis, J. R. (1994) Sample Sizes for Usability Studies: Additional Considerations. *Human Factors*, 36(2), pp. 368 – 378

Luff, P. and Heath, C. (1998) Mobility in Collaboration. In *Proceedings of CSCW'98.* ACM, pp. 305 - 314

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

Nielsen, C. (1998) Testing in the Field. In *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

Nielsen, J. and Molich, R. (1990) Heuristic evaluation of user interfaces. In Proceedings of CHI'90, Seattle, USA, ACM Press, pp. 249 – 256

Nielsen, M. (2000) Letting go the lines: Departure from the Felixstowe harbor. CHMI Research report CHMI-4-2000 http://www.cs.auc.dk/~pba/ElasticSystems