Studying Usability in Sitro

Kjeldskov, Jesper; Skov, Mikael

Published in:
International Journal of Human-Computer Interaction

DOI (link to publication from Publisher):
10.1080/10447310709336953

Publication date:
2007

Document Version
Peer reviewed version

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *22*(1 & 2), 7–36

Studying Usability In Sitro: Simulating Real World Phenomena in Controlled Environments

Jesper Kjeldskov
Mikael B. Skov
Department of Computer Science, Aalborg University, Denmark

Increased complexity of organizations and emerging technologies poses new and difficult challenges for the evaluation of software systems. Several years of research have proven that usability evaluations are invaluable tools for ensuring the quality of software technologies, but the increased complexity of technology requires new ways of understanding and evaluating the quality of software systems. This article explores limitations, challenges, and opportunities for studying mobile technologies "in use, in situ;" in laboratories (in vitro); and in controlled high-fidelity simulations of the real world. The latter condition is called *in sitro*. This report comes from 2 different case studies of evaluating the usability of mobile systems within these 3 different conditions. Results show that it is possible to recreate and simulate significant elements of intended future use situations in laboratory settings and thereby increase the level of realism and maintain a high level of control. In fact, the in sitro condition was able to identify most of the same usability problems as found in the other conditions. However, the in situ evaluation proved to provide a level of realism that is difficult to achieve in laboratory environments.

1. INTRODUCTION

As stated in the introduction to the 2005 In-Use, In-Situ: Extending Field Research Methods Workshop, "the increasing complexity of organizations and systems of communication, and the fast pace of technological change and adaptation, poses a challenge for researching the cognitive, social and cultural impact of technology that is in use in its natural settings, in situ" (Amaldi, Satinder, Fields, & Wong, 2005). One of the areas where this statement seems to be of particular importance is within the research field of mobile human–computer interaction (HCI) and systems design, where the emergence of new mobile, pervasive, and ubiquitous technologies continues to extend the scope of computer use in the workplace, home,

Correspondence should be addressed to Mikael B. Skov, Department of Computer Science, Aalborg University Fredrik Bajers Vej 7, DK–9220 Aalborg East, Denmark. E-mail: dubois@cs.aau.dk

and public, and consequently calls for research into people's use of technology going beyond our traditional laboratory approaches.

In the proceedings of the first workshop on Human–Computer Interaction for Mobile Devices in 1998, researchers and practitioners were encouraged to further investigate the criteria, methods, and data collection techniques for studying mobile system use (Johnson, 1998). Of specific concern to the development of such methods and techniques, it was speculated that traditional laboratory approaches would not adequately be able to simulate the context surrounding the use of mobile systems and that evaluation techniques and data collection methods such as think-aloud, video recording, or observations would be extremely difficult in natural settings—in situ. These concerns have since been confirmed through a number of studies, for example, Graham and Carter (1999); Pascoe, Ryan, and Morse (2000); Rantaten et al. (2002); Brewster (2002); Esbjörnsson, Juhlin, and Östergren (2003); and Kjeldskov and Stage (2004).

A number of different techniques have been suggested for studying technology in use, in situ such as workplace observations, contextual inquiries, interviews, focus groups, automatic logging of user actions, acting-out in context, and cultural and technology probes. Although such techniques provide valuable insights into actual use of software technologies, they are often rather limited in their ability to assess the specific qualities of the technologies in use and weak in their ability to identify design problems and inform redesign. Contrasting these methods, several years of HCI research have proven that usability evaluations are invaluable tools for measuring and improving the quality of software technologies, and hence usability engineering is today an established discipline within interaction design with widely acknowledged techniques and methods. With the emergence of mobile, pervasive, and ubiquitous technologies and the fast speed of technological change and adaptation that these technologies involve, the field of usability engineering is now faced with challenges such as lack of realism and real-world richness. In our view, this indicates an opportunity for combining the strengths of in situ empirical methods and usability evaluation techniques to overcome some of their individual shortcomings.

In 2003, a literature study on mobile HCI research methods revealed that 41% of the mobile HCI research and design reported in the main literature from 2000 to 2002 involved some sort of usability evaluation (Kjeldskov & Graham, 2003). However, even though evaluations of mobile systems are thus clearly prevalent, surprisingly little research had (and still has) been published concerning the methodological challenges just described. Exceptions include studies comparing two or more methods applied for evaluating mobile prototype systems in, for example, Brewster (2002); Graham and Carter (1999); and Pirhonen, Brewster, and Holguin (2002). Consequently, there is as yet no agreed set of appropriate usability evaluation methods and data collection techniques within the field of mobile HCI, and we still have little knowledge about the relative strengths and weaknesses of laboratory-based and field-based usability evaluations of mobile systems. Although the literature study (Kjeldskov & Graham, 2003) also revealed that 71% of mobile device evaluation was done through laboratory experiments and only 19% through field studies, it seems implicitly assumed that usability evaluations of mobile devices should be done in the field (Abowd & Mynatt, 2000; Brewster, 2002; Johnson,

1998). However, field-based usability studies are not easy to conduct. They are time consuming, and the added value is questionable. For discussions of some of the challenges of evaluating mobile systems in the field, see, for example, Pascoe et al. (2000), Rantanen et al. (2002), Esbjörnsson et al. (2003), and Kjeldskov and Stage (2004). Partly motivated by these challenges, some have suggested that instead of going into the field when evaluating the usability of mobile devices and services, adding mobility or other contextual features such as scenarios and context simulations to laboratory settings can contribute to the outcome of laboratory evaluations while maintaining the benefits of a controlled setting. For usability studies of mobile devices and services simulating mobility or other contextual factors in laboratory settings, see, for example, Salvucci (2001); Lai, Cheng, Green, and Tsimhoni (2001); Bohnenberger, Jameson, Krüger, and Butz (2002); Pirhonen et al. (2002); Kjeldskov and Skov (2003); and Kjeldskov and Stage (2004).

The purpose of this article is to contribute to the body of research on appropriate methods and techniques for evaluating mobile systems use by exploring the differences and similarities between studying such systems in use, in situ; in the laboratory; and in controlled high-fidelity simulations of the real world. We do this on the basis of two case studies of mobile system evaluation for real-world work tasks in highly challenging use contexts. These two case studies involve four empirical evaluations of mobile systems carried out in three different experimental conditions on the continuum from laboratory (in vitro) to field (in situ). On the basis of the two case studies, we outline limitations and challenges of evaluating mobile technologies in laboratory settings and in the real world. In response to these limitations and challenges, we have experimented with the use of a complementary approach, evaluating "in sitro," where real-world phenomena are simulated in a controlled environment. Based on a comparison of the usability evaluation results produced from each of these three conditions (in situ, in vitro, and in sitro), we explore the relative strengths and weaknesses of in sitro evaluations in comparison with in vitro and in situ evaluations.

The article is structured as follows. First we briefly highlight and discuss the value of studying technology use through the lens of usability. We then take up the discussion of trade-offs between realism and control when evaluating in laboratory (in vitro) and field settings (in situ), and we discuss the intermediate approach of evaluating in sitro. Based on this discussion, we map our two case studies of mobile systems evaluation onto a continuum of in situ, in sitro, and in vitro evaluation approaches outlining the relationships between four different empirical usability evaluations of mobile systems that we have carried out over the last 3 years. Sections 3 and 4 describe our two case studies of mobile systems usability evaluation. Section 3 describes a comparative usability study of a mobile system for communication onboard large container vessels, where we took up the challenge of increasing laboratory realism. We present the context for the study, the system developed, and two evaluations carried out in a traditional laboratory setting and in a high-fidelity ship simulator. Following this, we outline and compare the findings from these two evaluation approaches. Section 4 describes a comparative usability study of a mobile electronic patient record (EPR) system for use in a hospital ward where we took up the challenge of going into the field for the purpose of evaluating the system's usability. Again, we present the context for this study, the system devel-

oped, and two evaluations carried out in a simulated hospital ward and at a hospital during real work activities. We also outline and compare the findings from these two evaluation approaches. In section 5, we take a step back and highlight and explore the differences and similarities between our three experimental approaches, and we discuss the implications of our findings in relation to the issue of evaluating technology in use in situ. Finally, section 6 concludes our research and points out avenues for further work.

2. EVALUATING THE USABILITY OF MOBILE SYSTEMS

Several years of research have proven that usability evaluations are invaluable tools for ensuring the quality of software technologies. Therefore, usability evaluation of stationary computer systems is an established discipline within HCI with widely acknowledged techniques and methods. Several well-known textbooks on usability testing and engineering describe and illustrate how to plan, design, and conduct evaluations (e.g., Dumas & Redish, 1999; J. Nielsen, 1993; Rubin, 1994). These have contributed to improved evaluations and have had industrial impact. Furthermore, several attempts have "evaluated evaluations," that is, empirical evaluations of the relative strengths and weaknesses of the different approaches and techniques under different circumstances—for example, differences between think-aloud testing and heuristic evaluations testing (Bailey, Allan, & Raiello, 1992; Karat, Campbell, & Fiegel, 1992) and different user-based evaluation methods (Henderson, Podd, Smith, & Varela-Alvarez, 1995; Molich et al., 1998). So far, these kinds of comparative studies are only beginning to emerge in relation to the evaluation of mobile computer systems. A significant proportion of mobile technologies take many of the well-known methodological challenges of evaluating usability to an extreme. Users are often ambulatory, typically highly mobile during their interaction with the system, and situated in a dynamic and sometimes unknown use setting (Vetere, Howard, Pedell, & Balbo, 2003). The information presented to the users of mobile systems is closely related to their physical location, to objects in their immediate surroundings, or to their present as well as planned activities (e.g., Chincholle, Goldstein, Nyberg, & Erikson, 2002, Pospischil, Umlauft, & Michlmayr, 2002). Such challenges raise a number of interesting issues to consider when trying to understand the usefulness and usability of mobile systems. In particular, several discussions have been raised to determine when to evaluate mobile systems in vitro, as in laboratories, and when to evaluate mobile systems in situ, as in real-use context (Kjeldskov & Graham, 2003).

2.1. In Situ or In Vitro: The Trade-Offs Between Realism and Control

In situ and in vitro evaluations inherently integrate a number of characteristics.

- *In situ*, "in its original place": This condition defines that it is in its original location. For experiments involving the evaluation of computer systems, this often means that the use of the system takes place in its natural environment.

- *In vitro*, "in glass": This condition is distinguished from conditions that actually apply in nature. For experiments involving the evaluation of computer systems, this often refers to experiments that take place in controlled environments, such as usability laboratories.

In situ experiments are often characterized by a high level of realism (as illustrated in Figure 1). When dealing with evaluations of software systems, in situ experiments involve real users interacting with the system in a real situation and in the real context of intended use. Thus, the empirical basis for assessing the quality of the system is often very realistic. On the other hand, in situ evaluations are not easy to conduct (Brewster, 2002) and applying established evaluation techniques and data collection instrumentation, such as multicamera video recording, think-aloud protocols, or shadowing may be difficult in natural settings (Sawhney & Schmandt, 2000). Further, in situ evaluations complicate data collection because users are moving physically in an environment over which we have little control (Johnson, 1998; Petrie, Johnson, Furner, & Strothotte, 1998) and only partially comprehend. Also, for several mobile systems it is difficult to define and describe the original location, as location can be distributed in both time and space. Finally, some in situ evaluations may be impossible to conduct due to ethics or safety-critical issues (Kjeldskov & Skov, 2003).

In vitro experiments, on the other hand, are often characterized by a high level of control (as illustrated in Figure 1). For interaction design or HCI, in vitro evaluations often refer to experiments that take place in controlled environments, such as usability laboratories. In vitro evaluations can often benefit from experimental control and high-quality data collection when conducted in usability laboratories. Yet traditional usability laboratory setups may not adequately simulate the context surrounding the use of mobile systems. Thus, in vitro 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 re-create realistically in a usability laboratory. Second, working with systems for highly specific domains (Kjeldskov & Skov, 2003; Luff & Heath, 1998), laboratory studies may be impeded by limited access to prospective users on which such studies rely. Although benefitting 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.

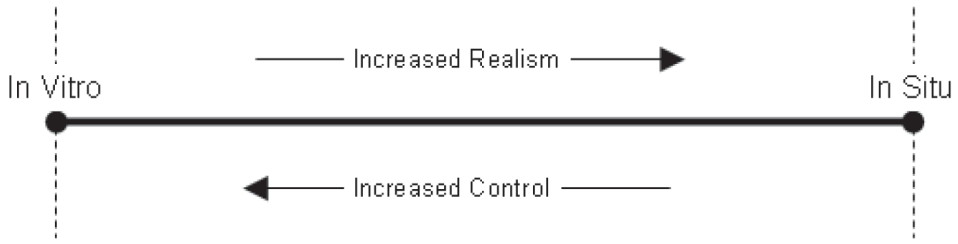

FIGURE 1 Simplified illustration of the often claimed trade-offs between a high level of control and a high level of realism in in situ and in vitro evaluations.

2.2. Simulation: An Attempt to Bridge Realism and Control

The inherent challenges of in situ and in vitro experiments related to realism and control have facilitated the introduction of additional experimental conditions. It is quite obvious that several of the outlined challenges cannot be solved through simple means as they are inherently integrated into the nature of the experiments, for example, the lack of realism when using a computer system in a laboratory. As a consequence, such challenges are often rather difficult to address and solve. However, a number of attempts have been suggested to overcome some of these difficulties. One viable way is simulation where selected elements of the experimental condition are simulated using computers or simulators. Several different types of simulations have been proposed and assessed, and different terms are often used for these simulations. In the following, we discuss two related but different types of simulations.

The first type of simulation is often referred to as *computational simulation* where computers fully simulate parts of an environment. Profoundly used in biology, such simulations serve to explore or investigate issues that are often difficult to do in vivo or in vitro; for example, Roulet et al. (1998) stated that computational molecular biology tools are becoming the method of choice for screening of certain DNA sequences. Computational simulations in biology have been coined *in silico* experiments (Wingender, 1998). This experimental condition stems from the Latin phrases *in vivo* and *in vitro*, which are commonly used in biology and refer to experiments done in living organisms and outside of living organisms, respectively. Wingender stated that in silico has been introduced into life sciences as a pendant to "in vivo" (in the living system) and "in vitro" (in the test tube) and implies the gain of insights by theoretical considerations, simulations, and experiments conducted on a silicon-based computer technology. Thus, simulation of real-world phenomena is important for such experiments. In silico experiments have further been adapted in other disciplines, for example, in computer science where Zhao, Stevens, Wroe, Greenwood, and Goble (2004) applied in silico experiments to simulate certain network behaviors. In summary, in silico experiments or computational simulations prove valuable when trying to understand effects of introducing new elements into a known environment that can be described (simulated) on a computer.

Although computational simulations provide promising conditions for experiments that can be fully automated, we bring attention to another kind of simulation referred to as simulators in which advanced high-fidelity, tailored environments provide a realistic context for human activity, for example, training, system design, or personal assessment (Sanders, 1991). For this particular stream of simulation, it is a technique substituting a synthetic environment for a real one, so that it is possible to work under laboratory conditions of control (Harman, 1961). Hence, experimenters are able to obtain a significant high level of realism while maintaining control over the experimental condition. Simulations with simulators are widely adopted within ergonomics and human factors for primarily training purposes and secondarily system design and personal assessment purposes (Sanders, 1991).

Simulators provide a very useful experimental approach, but studies have stressed potential challenges characterizing their use. One key problem with simu-

lators is performance measurement. Vreuls and Obermayer (1985) found that system performance measurements in highly sophisticated simulators are virtually useless due to poor system design; part of the problem resides in the fact that it is not clear what should or could be measured. Another important issue to consider is validation. Sanders (1991) argued that validation of simulators is crucial to establish how well the simulation actually reflects reality. Not until the simulation has been satisfactorily validated can it be used itself to evaluate the effect of deviation from the full physical fidelity. Alexander, Brunyé, Sidman, and Weil (2005) also acknowledged the importance of fidelity and described it as the extent to which the virtual environment emulates the real world. Different subcategories of fidelity have been proposed, like physical, functional, cognitive fidelities (Allen et al., 1986; Hays & Singer, 1989) and psychological fidelity (Mayer & Volanth, 1985) where, for example, functional fidelity has been defined as the degree to which the simulation acts like operational equipment in reacting to the tasks executed by the trainee (Allen, Hays, & Buffordi, 1986). Highly sophisticated simulators can almost truly simulate the different subcategories of fidelity, but they are often rather expensive and not very lightweight (Alexander et al., 2005). So far, a lot of effort has been put into making the simulator as realistic as possible, and Sanders (1991) stated that full simulations should be a final test and demonstration of the suitability of a new design rather than an open-ended trail.

2.3. In Sitro: Striving for Mobile Usability Realism and Control

In this article, we take a slightly different approach to simulation compared to the types just illustrated when trying to evaluate the usability of mobile systems. We are also concerned with simulating real-world phenomena when trying to enhance a controlled laboratory setting, but none of the aforementioned outlined approaches for simulation fits our work properly. First, the in silico experiments require that the simulation is conducted on a computer (e.g., Wingender, 1998). This is not the case for the evaluations we are interested in as we explore human activities with computer artifacts. Human activity is central in simulators, but full simulations, as illustrated by Sanders (1991) and Hays and Singer (1989), tend to focus several aspects of the human activity and related challenges of measuring and tailoring the realism. For the usability evaluation, we are primarily concerned with the identification of usability problems that prohibit a successful and fruitful interaction with the mobile system. Therefore, we wish to create an environment that partly or fully simulates other activities found in the real-use context.

As a consequence, we coin an analogous term for conditions simulating real-world phenomena in controlled environments when evaluating computer systems: *in sitro*. In sitro is concatenated from *in situ* and *in vitro* and stresses the combinational nature of the two conditions and of simulation of context. We define it as follows:

> *In sitro*: "in simulated context." This experimental condition describes a partially or fully simulated controlled laboratory-based evaluation where the intended future *in use* situation is being simulated.

The principle idea behind in sitro experiments is that part of the real-world phenomena is simulated in the laboratory. As illustrated in Figure 2, the aim of in sitro experiments is to increase the realism of in vitro evaluations while increasing the level of control of in situ evaluations.

2.4. In Sitro: Empirical Investigation

We present two independent cases involving four studies of usability evaluations of mobile systems involving 24 participants. These two cases serve to illustrate opportunities and limitations of our proposed experimental condition, in sitro. Our empirical investigation of the in sitro condition is illustrated in Figure 3.

The investigation contains two cases (A and B) of evaluating usability of mobile devices; both cases contain two studies adopting different evaluation conditions. Case A focused on increasing laboratory realism for the evaluation of a mobile system for coordination and collaboration on a large container vessel contrasting the use of a traditional laboratory setup (in vitro) with a high-fidelity simulation of the intended use context (in sitro). Case B focused on a mobile system for health care contrasting the use of a high-fidelity simulation of the use context (in sitro) with going out into the real world (in situ). The two cases are presented and discussed in sections 3 and 4.

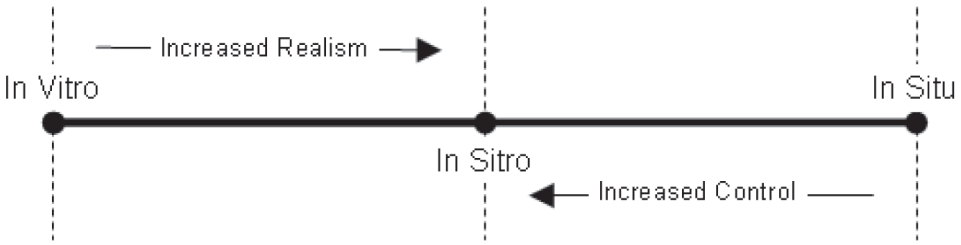

FIGURE 2 In sitro evaluations with increased levels of control and realism.

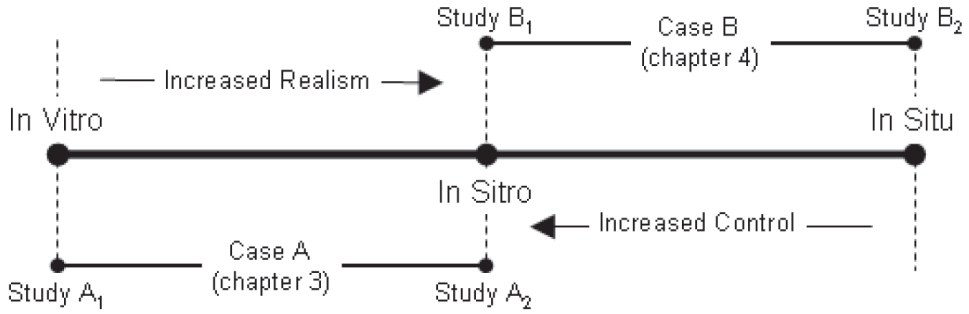

FIGURE 3 Illustration of our two studies that investigate the opportunities and limitations of in sitro experiments.

3. INCREASING LABORATORY REALISM

Our first comparative usability case (Case A) focused on the opportunities and challenges of increasing laboratory realism for the evaluation of a mobile system contrasting the use of a traditional laboratory setup (in vitro) with a high-fidelity simulation of the intended use context (in sitro; Kjeldskov & Skov, 2003). This study originated from our involvement in a large multidisciplinary research project involving ethnographic field studies of work activities in the maritime domain involving computerized process control and information systems (Andersen, 2000; M. Nielsen, 2000). As a part of this project, a mobile communication and coordination system, the Maritime Communicator, was developed for workers performing safety-critical collaborative work tasks onboard very large container vessels. Evaluating the usability of this system was a particular challenge for several reasons. First, the evaluation could not be done in situ for safety reasons but had to be done without going onboard the container vessels. Second, the use of the system was closely related to highly contextualized work activities in a very specialized physical use context, which would be difficult to recreate realistically in vitro. Motivated by these challenges, we decided to explore a series of different opportunities for increasing evaluation realism in controllable and safe environments.

We briefly present the Maritime Communicator case study next and describe how the two evaluation studies were designed and carried out.

3.1. Case A: The Maritime Communicator

The Maritime Communicator system was developed for supporting work activities onboard large container vessels (with sizes equivalent to three and a half soccer fields). The operation of such vessels requires workers to be highly mobile and physically distributed. Typically, the number of crew members is low, and hence people are assigned to various tasks at different locations on the ship depending on the situation: cruising at sea, departing from the quay, and so on. 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 when erroneous actions can cause serious material damage and possible injuries on personnel or loss of human life. Thus, systems for supporting these work activities must be carefully evaluated.

Distributed work activities in the maritime domain. On the basis of ethnographic studies of work activities on a container vessel (M. Nielsen, 2000) the Maritime Communicator was developed to support the coordination of "letting go the lines" immediately before departing from a harbor. When departing from a harbor the first step is to let 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.

Due to the enormity of the container vessel and the risk of lines getting sucked in and wrapped around the propeller or thrusters, leaving the vessel without any means of steering, the work tasks involved are distributed among a number of actors

located at strategic positions (Figure 4). These actors are all highly mobile throughout the whole operation. On the bridge (1), chief officers control the rudder, propeller, and thrusters. At fore (2) and aft (3), 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 of the operation consists of bringing the vessel clear of the quay sideways without running aground in shallow water or colliding with other ships. Because of wind, current, temporal lack of propulsion while lines are in the water, and poor visual view from the bridge, the operation of letting go the lines is not trivial and relies heavily on ongoing communication and careful coordination.

At present this coordination is primarily based on oral communication following a set of formalized procedures. Although people on the bridge can see and hear each other, personnel on deck are out of direct visual and audio contact and must communicate with the captain via walkie-talkies. To carry out the operation of departure, the captain needs an overview and total control over the propulsion, direction, and mooring of the ship. Although information about the rudder, propeller, and thrusters is 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 because disparity between the captain's mental model and the real world may cause wrong decisions, considerable cognitive resources are spent on establishing and maintaining common ground among the cooperating actors (Clark & Schaefer, 1989). Though flexible, radio-based communication suffers from limitations of technology as well as spoken language itself. Sound quality is often poor, utterances are not persistent, and communication is time consuming and suffers from language barriers and bottlenecks (multiple parallel tracks). Furthermore, it cannot be automated or integrated with other systems.

The prototype system. Inspired by the potentials of text-based messaging as an asynchronous, flexible, ubiquitous, and persistent communication channel requiring low cognitive overhead (see, e.g., Churchill & Bly, 1999), it was the thesis of the research team that a text-based communication channel on mobile devices could eliminate or reduce some of limitations observed during the field studies. To investigate this potential further, a prototype of the Maritime Communicator was

FIGURE 4 Sine Maersk in Gothenburg container terminal. 1 = bridge, 2 = fore area, 3 = aft area.

designed and implemented (Kjeldskov & Stage, 2003; see Figure 5). The prototype setup consisted of three iPAQ 3630 connected through an IEEE 802.11b 11Mbit wireless TCP/IP network. One device was intended for the captain on the bridge, and the other two were intended for first and second officers on the fore and aft deck, respectively. The Maritime Communicator 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.

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 overall operation and commands appear only 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. Above this list, the workers can monitor ongoing threads of communication as they unfold during the operation synchronized with the graphical representation of the vessel and mooring lines.

In the following sections, we describe two evaluations of the Maritime Communicator carried out in vitro and in sitro.

3.2. Study A_1: Laboratory Evaluation (In Vitro)

In our first evaluation study, we focused on evaluating the usability of the Maritime Communicator in vitro: through a "traditional" laboratory-based think-aloud evaluation with prospective users as described by, for example, Rubin (1994). In vitro evaluations are not by definition nonrealistic just because they do not take place in the intended situation of use. In our first study, for example, some realism was provided through (a) the tasks to be solved using the system; (b) the physical separation of the communicating test participants; and (c) a simple cardboard mock-up of the vessel, quay, and mooring lines.

The first study was conducted in a standard usability laboratory consisting of two separate participant rooms (resembling the bridge and the fore deck, respectively) and a control room. From the control room, both participant rooms could be surveyed through one-way mirrors and by means of remote-controlled motorized cameras mounted in the ceiling. Six test participants took part in the study. They

FIGURE 5 The Maritime Communicator.

were divided into three teams of two and given the task of letting go the lines before departure of a large vessel coordinating the operation by means of the Maritime Communicator. All test participants 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. They were recruited from the nearby Skagen Maritime College. The test participants received a 15-min joint oral introduction to the specific 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 6). The introduction and use scenario covered the overall operation of letting go the lines, the basic concepts and maritime notions involved, the distribution of work tasks, and present procedures of communication and coordination (as just described). Following this, one person was asked to take the role of captain on the bridge, and the other took the role of officer on the fore mooring deck.

The test participants were seated at a desk with the mobile device located in front of them. During the evaluation, the test participants were asked to think-aloud, explaining their comprehension of and interaction with the prototype. Supporting this, the captains were given a cardboard mock-up of central instruments on the bridge for controlling the thrusters, propellers, and rudder as well as a model of the ship and mooring lines placed on a schematic drawing of the harbor (Figure 7). The purpose of this mock-up was to supply the test participants with a tool for explaining and illustrating their strategies and actions as the process of departing from the harbor developed over time. An evaluator located in each test room observed the test participants and frequently asked them about their actions. On a video monitor facing away from the test participants, the evaluators could see a close-up view of the mobile device as well as the activities in the other participant room for the sake of overview. The evaluations lasted approximately 30 min and were followed by a 10-min debriefing interview.

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 using a predefined script. Two A4

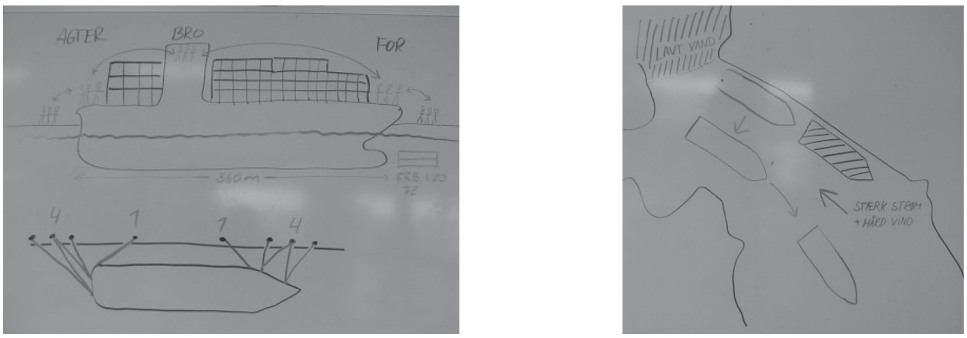

FIGURE 6 Introduction to use context and a possible use scenario drawn on whiteboard.

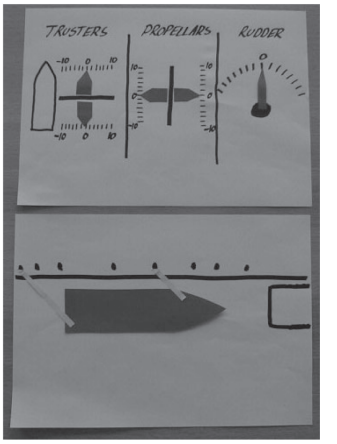

FIGURE 7 Cardboard mock-up of the bridge, ship, and mooring.

handouts depicted standard patterns of mooring and explained 10 basic concepts and notions of the maritime context for quick reference if necessary.

Remote-controlled video cameras mounted in the ceiling captured high-quality video images of the evaluation sessions. Two cameras captured overall views of the captains and officers, and two cameras captured close-up views of the mobile devices. To ensure good video images of the displays, the test participants 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 8). Audio from the two participant rooms was recorded on separate tracks for later mixing and potential separation during analysis.

3.3. Study A_2: Simulating the Ship (In Sitro)

In our second evaluation study, we aimed at evaluating the Maritime Communicator prototype in the hands of real users in a highly realistic but yet controllable and safe environment, thus combining strengths and benefits from both in situ and in vitro studies. We define this approach as *in sitro*. Accomplishing this aim, we established a temporary usability laboratory at the simulation division of Svendborg In-

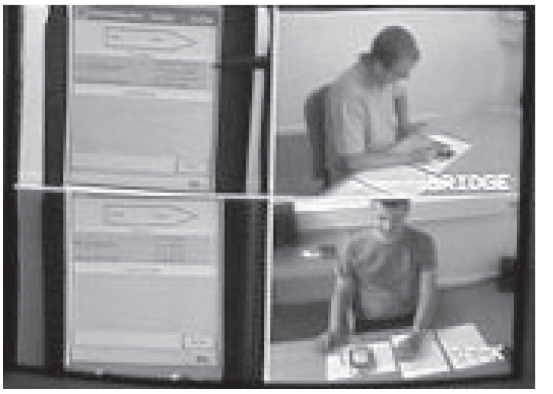

FIGURE 8 Video recording from evaluation in the usability laboratory.

ternational Maritime Academy and used their state-of-the art ship simulator for creating a realistic (but safe) setup simulating real-world phenomena from the intended use context on a high level of fidelity. The ship simulator consisted of two separate rooms: a simulated bridge (see Figure 9) and a nearby control room. The bridge was fully equipped with controls for thrusters, propellers, rudder, and so on, as well as instruments, such as dobler log, echo sounder, electronic maps, radars, and VHF radio. 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. Weather and dynamic traffic conditions also could be included into the scenario. For our specific study, 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 (M. Nielsen, 2000).

As in our first study, three captains and three officers, divided into teams of two, participated as test participants in the study fulfilling their usual roles and were given the overall task of letting go the lines and departing from harbor using the Maritime Communicator for communication between bridge and deck. Again, all participants were educated and practically experienced prospective users fulfilling their usual roles in the use domain—this time recruited from the academy running the simulator facility. Carrying out the operation, the captain had to consider all aspects of maneuvering the ship on the simulated bridge. This included controlling the rudder, propellers, and thrusters as well as communicating with personnel on the ship, harbor traffic control, and so on, and taking into consideration the movements of other vessels. The primary task of the first officer on deck (located in the neighboring simulator control room) was to orally forward commands executed by the captain via the mobile device prototype to the operator of the simulation (impersonating 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 (making the vessel respond differently to controls on the bridge as it would in the real world) and report to the first officer when the requested operations (such as letting go a line) had been carried out. For simplicity, commands targeted at the second officer on the aft deck were fed directly into the simulation, and the simulation operator gave feedback.

FIGURE 9 The part of the simulated bridge at Svendborg International Maritime Academy.

During the evaluation, the captain and officer 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 participants and asked questions for clarification. On a video monitor facing away from the test participant, the evaluator on the deck could see a close-up view of the mobile devices and an overview of the bridge (see Figure 10). The evaluations lasted approximately 40 min and were followed by a 10-min debriefing interview.

As in the traditional laboratory study, 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, and close-up views of the test participant's interaction with the prototype and other controls on the bridge were captured by the evaluator using a handheld camera. In the room resembling the fore mooring deck, a camera captured an overall view of the test participant and the operators of the simulator. As in the traditional usability laboratory, the test participant acting as the 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 to ensure good video images of the display. The four video signals were merged into one composite signal and recorded digitally. Audio from the two rooms was recorded on separate audio tracks.

3.4. Analysis

The data analysis from Studies 1 and 2 aimed at identifying, describing, and classifying usability problems experienced during use of the Maritime Communicator prototype. The analysis of the data from the two studies was done in a collaborative effort between the two authors, allowing immediate discussions of identified problems, and involved two discrete steps: a compilation of findings for each study and a comparison of findings across studies. To ensure a rigorous and credible process, we went through the following steps. First, problems experienced by the test

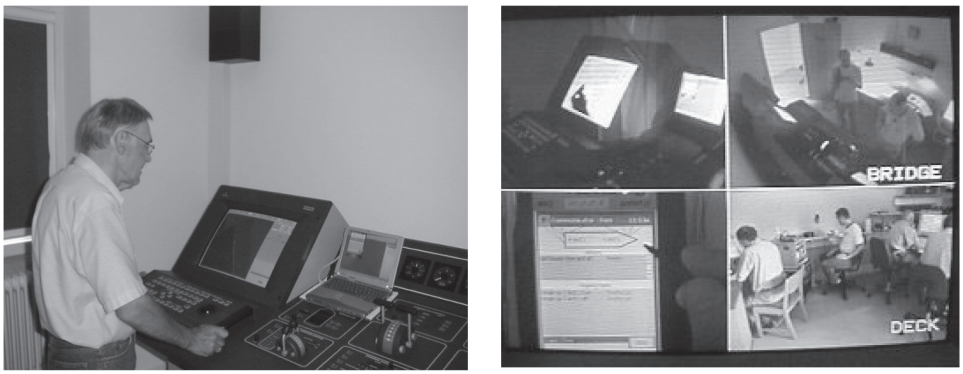

FIGURE 10 Studying use in the ship simulator: the simulator and the video recording.

participant acting as first officer were identified by examining the video recordings and only listening to the audio track from the fore deck. Second, problems experienced by the test participant acting as captain were identified examining the video recordings and only listening to audio from the bridge. Third, all video recordings were examined again while listening to a mix of the audio from both the fore deck and the bridge to "get the whole picture"—confirming the problems already identified and identifying additional problems. The compilation process resulted in two lists of usability problems ranked as cosmetic, serious, or critical (Molich, 2000).

Finally, the two lists of usability problems were merged into one complete list through extended discussion of each identified problem (member checking) among the authors until consensus had been reached. In case of different severity ratings of the same usability issue across techniques, the most severe rating was used in the merged list.

3.5. Findings

We identified 53 different usability problems from the six in vitro and in sitro sessions. Eight problems were critical, 20 were serious, and 25 were cosmetic. The in vitro sessions identified 40 of the 53 problems, whereas the in sitro showed 36 of the 53 problems. Twelve of the problems were unique to the in sitro sessions, whereas 17 problems were unique to the intro sessions. Most of the problems were experienced by many participants. Some of the problems were interaction issues; for example, nearly all test participants had problems about which elements to interact with on the screen, whereas a few related to the correlation between the representation of the ship on the system and real activities on the ship. As another example, many test participants could not state the status of commands they had issued.

Figure 11 outlines the distribution of the identified 53 usability problems; each column represents 1 usability problem associated with the number of test participants experiencing the problem (indicated by black boxes) for both settings. The distribution of problems on severity furthermore reveals that both conditions identified a large proportion of the critical and serious problems. However, the in sitro condition was able to reveal all 8 critical problems, whereas the in vitro condition only identified 6 of the 8 critical problems.

As just mentioned, 12 usability problems were unique to the in sitro condition, which constitute one third of the total identified problems for that condition. These

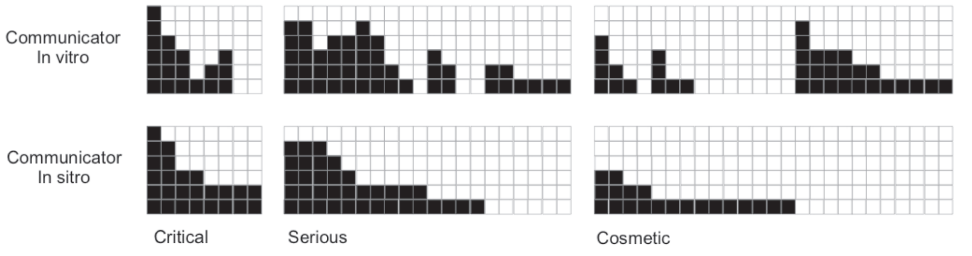

FIGURE 11 Distribution of identified usability problems.

problems primarily concern the representation of the task in the system and lack of flexibility, for example, more of the domain participants 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; for example, one test participant lost complete overview of what was going on because he had to cancel one action. Finally, it should be noticed that both conditions identified several unique cosmetic problems only experienced by one participant.

4. GOING INTO THE FIELD

Fueled by the challenges encountered and lessons learned from the comparative usability study of the Maritime Communicator, our second comparative usability case (Case B) focused on contrasting the use of a high-fidelity simulation of the use context (in sitro) with going out into the real world (in situ). Again, the motivation for doing so originated from our involvement in a larger research project—this time dealing with the use of computerized information systems in the health care domain with particular focus on the use of EPR systems at hospitals. As a part of this project, we developed MobileWARD—a mobile counterpart to the stationary EPR system of a large regional hospital supporting nurses' everyday collaborative and highly mobile work activities on a hospital ward. Although in this project studying the use of the prototype system in situ was not ruled out, we were still faced with a series of challenges similar to those of the Maritime Communicator study. First, hospital staff raised concerns that using the prototype EPR system could influence them in their taking care of their patients and potentially impact on their well-being. Second, significant ethical considerations were raised about involving real hospitalized patients (and potentially sensitive patient data) in a research study at all. Third, allowing researchers access to the hospital ward during hectic work hours (which was when the system was intended to be used) was not popular among the nurses who were already very busy and under considerable work pressure due to understaffing. Motivated by these challenges, our experience with the use of simulated use contexts in laboratory settings, and the ongoing discussion about laboratory versus field evaluations of mobile usability, we decided to carry out a comparison between the results produced when simulating real-world phenomena in a controlled environment (in sitro) and when studying usability purely in situ: observing real users doing real work in the real use context—with no researcher control or interference (to which the hospital staff eventually agreed).

Next we briefly present the MobileWARD case study and describe how the two evaluation studies were designed and carried out.

4.1. Case B: MobileWARD

MobileWARD was developed for supporting collaborative mobile work activities at a hospital ward through wireless access to electronic patient data on handheld computer terminals. Supporting work activities in health care is highly complex

and challenging, and within the last 20 years considerable effort has been devoted to the development of computerized systems for this domain, such as electronic patient records. An electronic patient record is a collection of information about a single patient's history in a hospital, which the hospital staff use to diagnose diseases and to document and coordinate treatment. The primary motivation for this effort is that unlike paper-based patient records, electronic patient records will be accessible to all relevant persons independent of time and location. The design of electronic patient records is a huge challenge for the HCI community, raising a wide range of still-unanswered questions related to issues such as screen layout, interaction design, and integration into work processes. However, although much research has studied the use of traditional paper-based patient records, suggesting electronic counterparts, little research has been published on studies inquiring into the use of the mass of EPR systems already out there.

Using electronic patient records in health care. Based on evaluations of EPR systems and field studies of mobile work activities in hospitals, we identified three key issues concerning the use of electronic patient records: mobility, complexity, and relation to work activities.

- *Mobility*. Most nurses expressed concerns about having to be mobile while working with the EPR system, which was stationary. Meeting this challenge, the use of laptop computers rather than desktop workstations had been suggested and discussed at the hospital. However, most of the nurses stated that they would find it impossible or unfeasible to carry a laptop computer around the ward every time they were to conduct work tasks away from their office. One problem was the size of the laptop, as they would also have to carry other instruments.
- *Complexity*. Another overall concern reported and observed was the complexity and fragmentation of information. Most nurses found it difficult to locate the necessary patient information in the EPR system to carry out their work. This sometimes led to delays and incomplete task completions. Hence, the nurses would be unsure whether they had found the right information and whether they had succeeded in finding all relevant information.
- *Work relation*. Most nurses experienced problems with the EPR system due to difficulties with relating the data and structure of information in the system to real work activities and people. The problem was that they would typically use different kinds of information in context to determine how to solve a problem—for example, the visible condition of a patient. Another concern related to the fact that the system only partially reflected their current work tasks, making it difficult to the test participants to find or store information.

The prototype system. Inspired by the potentials of context-aware mobile computing, it was our thesis that providing nurses with mobile access to electronic patient records automatically adapting to their current work situation could help overcoming some of the observed limitations of the stationary EPR system. To investigate this potential, a functional context-aware prototype system was designed

and implemented to support the nurses' morning procedure (Skov & Høegh, 2005), which (a) supported the highly mobile work activities of nurses by being handheld, (b) reduced complexity by adapting to its context, and (c) eliminated double registering of information (first written down on paper and later entered into the PC) by being integrated with the existing patient record. Facilitating access to patient information at the "point of care" is not a new idea (Arshad, Mascolo, & Mellor, 2003; Kaplan & Fitzpatrick, 1997; Urban & Kunath, 2002), but adapting information and functionality in a mobile EPR system to its context is a novel approach to improving the usability of such systems, which has not yet been investigated thoroughly.

MobileWARD runs on Microsoft PocketPC-based Compaq iPAQ 3630 (or equivalents) connected to an IEEE 802.11b wireless TCP/IP network. In the intended setup, all nurses on duty have their own personal device. The MobileWARD system is context aware in the sense that the system presents information and functionality adapted to the location of the nurse, the time of the day, pending tasks, nearby patients, and so on. Based on the classification by Barkhuus and Dey (2003), MobileWARD is an active context-aware system as it automatically presents information and adapts to its context. The system works as described next.

Before visiting assigned patients for morning procedure, nurses often want to get an overview of the specific information about each patient. As this typically takes place at the nurse's office or in the corridor, the system by default displays the overall patient list (Figure 12a). Patients assigned for morning procedure are shown with a white background, and the names of patients assigned to the nurse using the system are boldfaced (e.g., "Julie Madsen"). For each patient, the patient list provides information about previous tasks, upcoming tasks, and upcoming operations. The indicators TP (temperature), BT (blood pressure), and P (pulse) show the measurements that the nurse has to perform. "O" indicates an upcoming operation (within 24 hr), which usually requires that the patient should fast and be prepared for operation. At the top of the screen, the nurse can see his or her current physical location (e.g., "in the corridor").

The window in Figure 12b displays information related to one patient, including the patient's name and personal identification number, previous sets of measured temperatures, blood pressures, and pulses, as well as notes regarding the treatment

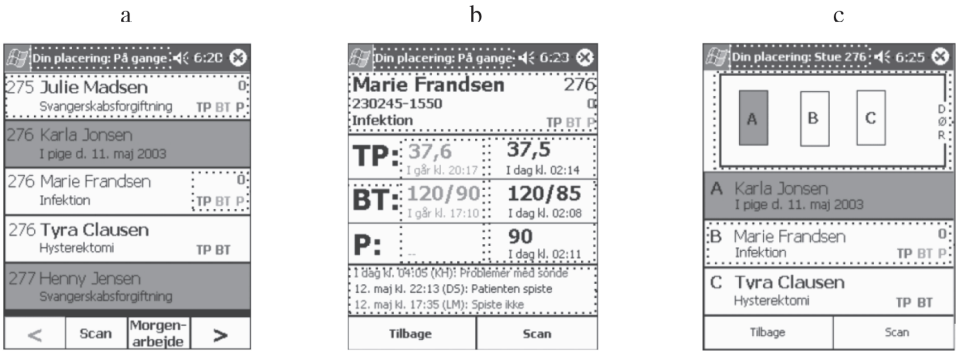

FIGURE 12 MobileWARD: Three different screens from the context-aware mobile EPR system.

of the patient. To enter new data into the system, the nurse must scan the bar code identification tag on the patient's wristband using the "Scan" function in the bottom of the screen. When the nurse enters a ward, the system automatically displays information and functionality relevant to this location (Figure 12c). Information about the patients on the current ward is presented, resembling the information available on the patient list displayed in the corridor, with the addition of a graphical representation of the physical location of the patients' respective beds. Data on each patient are available by clicking on the patient name.

In the evaluated prototype of MobileWARD, some of the contextual sensing functionality was simulated by means of a "context control center" application. The control center runs on a separate iPAQ connected to the wireless network. Through this application, an operator can trigger "context events" in MobileWARD, for example, instructing the system that the user has entered a specific room.

4.2. Study B_1: Simulating the Hospital Ward (In Sitro)

The aim of our third study was to evaluate MobileWARD in a controlled simulated environment (similar to the ship simulator) where we could closely monitor the use of the system and simulate key real-world phenomena such as mobility between rooms, work tasks, and hospitalized patients. To achieve this, we modified and refurnished our usability laboratory to resemble a part of the physical space of a hospital department (Figure 13). This included the use of two separate evaluation rooms connected by a hallway. Each of the evaluation rooms were furnished with beds and tables similar to real hospital wards. From a central control room, the evaluation rooms and the hallway could be observed through one-way mirrors and via remote-controlled motorized cameras mounted in the ceiling.

Six test participants (four women and two men) between 28 and 55 years of age took part in the study. All test participants were trained nurses employed at a large regional hospital and had between 2 and 36 years of professional experience. Thus

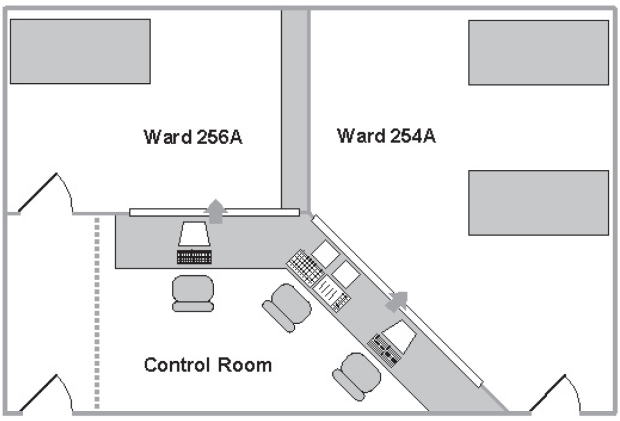

FIGURE 13 Physical layout of the usability laboratory simulating the hospital ward.

as in the maritime communicator evaluations, the test participants were real prospective users fulfilling their usual roles. All test participants were mobile phone users, but only one had experience with the use of handheld computers. Everyone was also familiar with stationary EPR systems and described themselves as experienced or semiexperienced information technology (IT) users. All test participants were given a series of tasks to solve while using the system. The tasks were derived from a field study at a hospital ward and were developed in collaboration with hospital staff. The tasks covered the duties involved in conducting standard morning work routines involving primarily (a) checking up on a number of assigned patients based on information in the system from the previous watch; (b) collecting and reporting scheduled measurements such as temperature, blood pressure, and pulse; and (c) reporting anything important for the ongoing treatment of the patients that should be taken into consideration on the next shift.

Before the evaluation sessions, the test participants were given a brief introduction to the system, including the room-sensing functionality and the procedure for scanning patients' bar code tags. The test participants were also instructed on how to operate the available instruments for measuring temperature, blood pressure, and pulse. The evaluation sessions were structured by the task assignments, which required the test participants to interact with all three patients in the two simulated hospital wards and to move between the two rooms through the connecting hallway a number of times. The nurses were encouraged to think-aloud throughout the evaluation, explaining their comprehension of and interaction with the system. The evaluations lasted between 20 and 40 min, after which the participants filled out a questionnaire.

Each evaluation session involved six people. One nurse used the system for carrying out the assigned tasks. Three students acted as hospitalized patients. One researcher acted as test monitor and asked questions for clarification. A second researcher operated the context-control center and the video equipment. For data collection, high-quality audio and video was recorded digitally from the ceiling-mounted cameras, and a tiny wireless camera clipped onto the mobile device provided us with a close-up view of the screen and of user interaction (Figures 14 and 15).

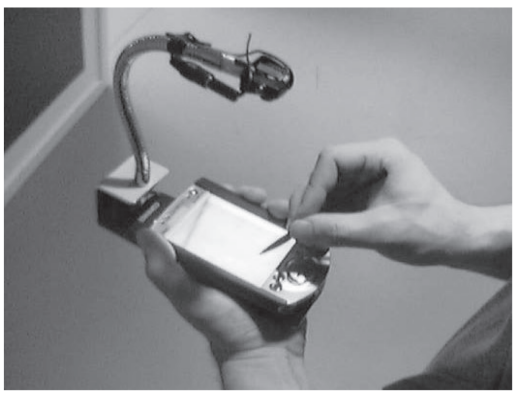

FIGURE 14 Wireless camera mounted on the personal digital assistant.

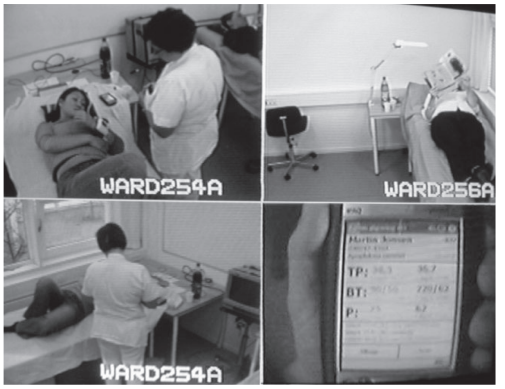

FIGURE 15 Video images from simulated ward and personal digital assistant.

4.3. Study B_2: Studying Use at the Hospital (In Situ)

The fourth evaluation study took place at a large regional hospital in Denmark. The aim of this evaluation was to study the usability of MobileWARD in situ for supporting real work activities at a hospital involving real nurses, real patients, and real patient data. To achieve this, we adopted an observational approach combined with questions for clarification while the nurses were not directly engaged in conducting their work.

The in situ evaluation was carried out at the Medical Department at the Hospital of Frederikshavn (Figure 16). This space included the physical area of seven hospital wards, an office with reception, a rinse room, and a living room connected by a central hallway, and the evaluation involved nurses at work and patients committed to the hospital. Six test participants (all women) between 25 and 55 years of age participated in the in situ evaluation. All test participants were trained nurses employed at the Hospital of Frederikshavn and had between 1 and 9 years of professional experience. They were all mobile phone users but novices with the use of handheld computers. All test participants were frequent users of a stationary EPR system and described themselves as experienced or semiexperienced users of IT.

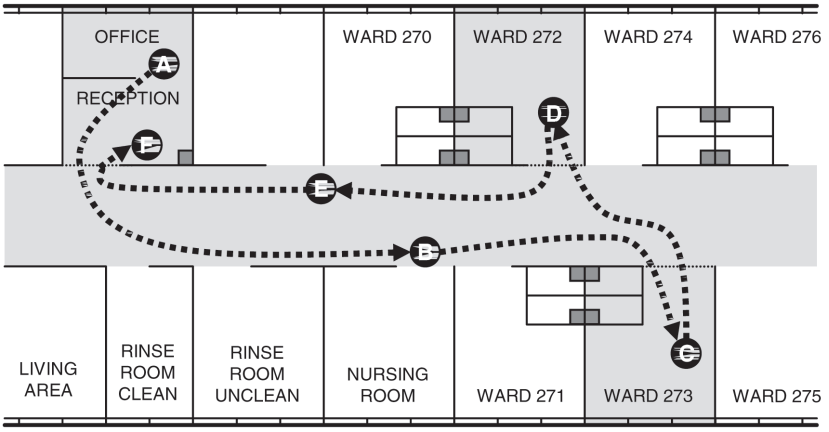

FIGURE 16 Physical layout of the hospital wards.

The in situ evaluation did not involve any researcher control and interference in form of task assignments but was structured exclusively by the work activities of the nurses in relation to conducting their standard morning work routines. As in the task assignments of the laboratory evaluation, this involved (a) checking up on a number of assigned patients in different wards and moving between different rooms through the connecting hallway a number of times, (b) collecting and reporting scheduled measurements, and (c) reporting anything important for the ongoing treatment of the patients. As in the laboratory evaluation, the test participants were given a brief instruction to the MobileWARD system, including the room-sensing functionality and the procedure for scanning a patient's bar code tag. The evaluations lasted 15 to 20 min on average and were followed by the completion of a brief questionnaire. To be able to include a suitable number of nurses, the study took place over 2 days.

Each evaluation session involved six people. One nurse used the system for carrying out her work activities. One researcher observed the work and use of the mobile system from a distance and asked questions for clarification while in the hallway. A second researcher operated the context-control center application and the portable audio/video equipment. In addition, each evaluation session involved three hospitalized patients in their beds. Due to the real-life nature of the study, each evaluation session involved different patients, and the nurses did not think-aloud.

Due to the challenges of capturing high-quality data during usability evaluations in natural settings (e.g., Brewster, 2002; Esbjörnsson et al., 2003; Kjeldskov & Stage, 2004; Pascoe et al., 2000; Rantanen et al., 2002), we designed and purpose-built a portable configuration of audio and video equipment to be carried by the test participant and an observer, allowing a physical distance of up to 10 m between the two. The configuration consisted of a tiny wireless camera (also used in the laboratory evaluation just described) clipped onto the mobile device (Figure 14) and a clip-on microphone worn by the test participant. Audio and video were transmitted by wireless to recording equipment carried by the observer (Figure 17).

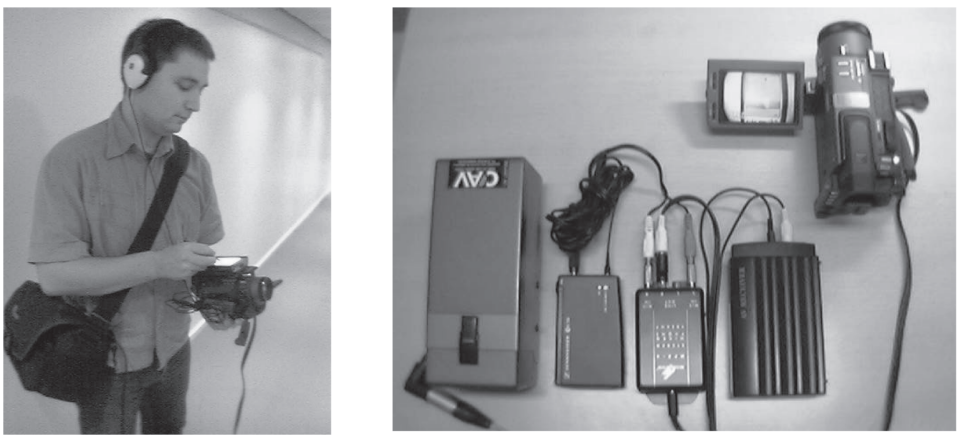

FIGURE 17 Observer (left) carrying and operating portable audio/video equipment (right) for capturing high-quality data in the field.

In the test monitor's bag, the video signal from the clip-on camera was merged with the video signal from a handheld camcorder (picture-in-picture) and recorded digitally. This setup allowed us to record a high-quality close-up view of the screen and user interaction as well as an overall view of user and context. During the evaluation, the observer viewed the user's interaction with the mobile device on a small LCD screen and monitored the sound through headphones. For ethical reasons, we were not permitted to film the hospitalized patients.

4.4. Analysis

The data from studies 3 and 4 amounted to approximately 6 hr of video recordings depicting the 12 test participants' use of the MobileWARD system. On the basis of these data, the analysis aimed at identifying, describing, and classifying two lists of usability problems experienced by the users in the two studies. All sessions were analyzed in random order by two teams of two trained usability researchers holding doctoral or master's degrees in HCI. As in the analysis of the Maritime Communicator studies, the data analysis involved a compilation of findings for each study and a comparison of findings across studies. Compiling the usability problems, each team first analyzed the videos recordings in a collaborative effort allowing immediate discussions of identified problems and their severity (cosmetic, serious, or critical). As a guideline for the collaborative analysis, each identified usability problem would be discussed until consensus had been reached. The two teams of researchers produced two lists of usability problems indicating for each problem if it was experiences in sitro, in situ, or both. Subsequently, these two lists were merged into one complete list through a process of comparing each problem. Again, this was done in a collaborative effort, discussing each problem and its severity until consensus had been reached. In case of different severity ratings across techniques, the most severe rating was used.

4.5. Findings

We identified 37 different usability problems from the 12 in sitro and in situ sessions. Eight problems were assessed to be critical, 19 problems were serious, and 10 were cosmetic. Our case showed that the in sitro condition found more usability problems than the in situ condition. The six in sitro participants experienced 36 of the 37 usability problems, whereas the six in situ participants experienced 23 of the 37 usability problems. Fourteen usability problems (1 critical, 9 serious, 4 cosmetic) were unique to the in vitro condition, whereas 1 serious usability problem was unique to the in situ condition.

Regarding the critical problems, the in sitro setting identified all eight critical problems and the in situ setting identified seven critical problems. Considering the serious problems, we find that the in sitro identified eight extra problems compared to the in situ evaluation.

Figure 18 outlines the distribution of the identified 37 usability problems where each column represents one usability problem associated the number of test participants experiencing the problem (indicated by black boxes) for both settings. Seven usability problems (2 critical, 2 serious, 3 cosmetic) were experienced by all 6 participants in the in sitro setting, whereas 3 usability problems (2 serious, 1 cosmetic) were experienced by all 6 participants in the in situ setting, and 1 usability problem (cosmetic) was experienced by all 12 participants.

Looking across the distribution of the usability problems, we find that although the critical problems have a roughly similar distribution, the serious and cosmetic problems have rather dissimilar distributions where some problems were identified by all or nearly all participants in one setting but only identified by a few or none in the other setting. For example, all participants were informed to use either their fingers or the attached pen for device interaction, but only the in sitro participants chose to use the pen, and most of them experienced difficulties in placing the pen between tasks.

One problem was identified by only in situ participants. This problem concerned the validity of recorded data in the system. Two of the nurses reported and recorded accurate data only on patients' heart rate, temperature, and blood pressure. Occasionally the nurses would first measure these values and then perform other work activities. Later, when recording the values into the system, they had forgotten the exact measures and would then repeat the measures. This was not the case in the in sitro condition where the participants stressed the artificial condition and the lack of need for being accurate. Apparently, the in use and in situ situation made the participants stress accuracy and validity.

5. DISCUSSION

Our motivation behind preparing this article was to contribute to the body of research on development of appropriate methods and techniques for evaluating mobile system use by systematically exploring the differences and similarities between studying such technologies "in use, in situ;" in the laboratory; and in controlled high-fidelity simulations of the real world. Thus, our aim was to contrib-

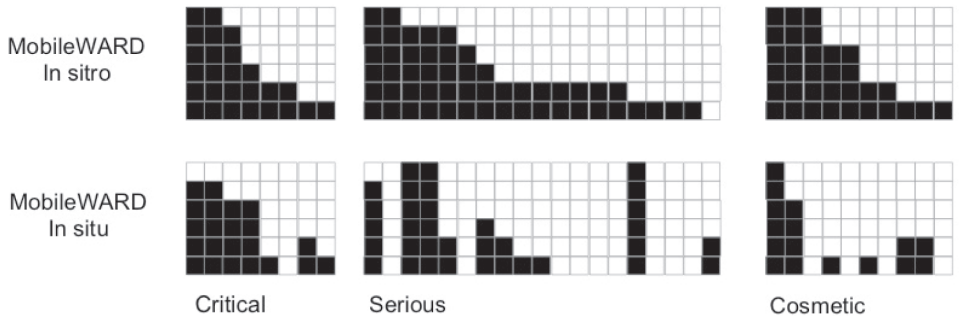

FIGURE 18 Distribution of identified usability problems.

ute to the general knowledge of usability evaluation and testing on how to set up an environment for testing (see, e.g., Dumas & Redish, 1999; J. Nielsen, 1993; Rubin, 1994) and more specifically to the extensive body of knowledge on comparative usability evaluations studies in general (see, e.g., Bailey et al., 1992; Karat et al., 1992; Henderson et al., 1995; Molich et al., 1998) and especially for mobile systems evaluations (Kjeldskov & Stage, 2004). Our comparative usability evaluation study explored three conditions for mobile usability evaluations.

We proposed an experimental condition for mobile usability evaluations called in sitro from the in situ and in vitro conditions and stressed the combinational nature of the two conditions and of simulation. The idea of in sitro is to simulate partial or full fidelity of a real intended use situation. In sitro is closely related to and takes inspiration from research on simulation (see, e.g., Allen et al., 1985; Hays & Singer, 1989; Mayer & Volanth, 1985; Sanders, 1991). Our results indicated that the in sitro condition was able to simulate significant elements of the intended future in use situation and thereby increase the level of realism in the evaluation and maintain a high level of control.

Both investigated cases involved rather complex and dynamic use situations: captain and crew members coordinating activities on a large container vessel during departure from harbor, and nurses doing morning procedures at a hospital ward. The former setup involved a rather expensive and sophisticated simulator used for training of future container vessel captains, but the latter setup was rather simple, taking place in our traditional usability laboratory including two wards, a number of beds, and some patients (student actors). Although the container vessel setup resembles full simulations as described in Sanders (1991), the hospital setup was rather lightweight and resembles some of the ideas for training using desktop games illustrated in Alexander et al. (2005).

Both conditions could potentially be difficult to simulate in a laboratory; however, our studies confirmed that for usability evaluations this is possible. Our first case exhibited similar distribution of identified usability problems for the critical problems and partially similar distribution for the serious problems. In fact, the in sitro condition identified additional critical problems in the interface not found in the in vitro condition. Both these problems could be traced to the increased level of realism in the use situation. At the same time, we experienced no significant problems with experimental control when we introduced use situation elements, for example, other artefacts, additional participants, and the participants still being able to think-aloud during the evaluation. Thus, it seemed possible to simulate parts of the environment in a controlled laboratory. Our second case confirmed this observation as the in sitro condition was able to identify nearly all the problems identified in situ. Several problems were further identified only in the in sitro condition, but most of these were due to lack of control in the in situ condition.

The real-world condition (in situ) integrated levels of realism that were not achieved in the in sitro condition which primarily concerned the validity of the recorded data in the system. Thus, in our case it seemed impossible to re-create all levels of fidelity in the simulations. Some in situ participants would only report and record accurate (and thus realistic) data on patients' heart rate, temperature, or blood pressure. This observation was seen when some nurses first measured values and then performed other work activities. The nurses would record the values in the sys-

tem later but had by that time forgotten the exact values and would therefore repeat the measures. This was not the case in the in sitro condition, where several participants stressed the artificial condition and the lack of need for being accurate. As a consequence, they occasionally entered data into the system that were incorrect or simply wrong. We argue that this aspect has to do with psychological fidelity as illustrated by Alexander et al. (2005). They stressed that many real-world environments evoke levels of stress and arousal that may not be directly replicable in virtual environments. This seemed to be confirmed in our study. Other studies have investigated the effects of adding contextually relevant stress to training paradigms (Driskell, Johnston, & Salas, 2001), but this was not attempted in our approach. Although both conditions lacked psychological fidelity, the container vessel setup did integrate the same kind of problem, probably due to an established seriousness when using the simulator. Thus, even though stress could have been low as the participants were not maneuvering a real container vessel, all of the participants acting as captain took the assignment very serious and never entered false data deliberately into the system.

Our in situ sessions with nurses confirmed the challenges of decreased control as none of the participants used the note-taking facility in our electronic patient record prototype. In the in situ study, we deliberately chose to give them no assignments, as this would possibly increase level of control (and thereby perhaps decrease level of realism). As a direct consequence, we identified no usability problems in the note-taking facility of the prototype from the in situ condition. Thus, control was definitely a challenge in our study.

The in vitro condition provided only a few additional findings when compared to the in sitro condition. The in vitro sessions in our first case were easier to plan and conduct. In using our own usability laboratory, we could more easily set up and conduct the evaluations. Comparatively, the in sitro sessions required more resources and more planning. Furthermore, the in vitro sessions identified a number of serious and cosmetic problems not identified in sitro. It is difficult to assess the value of these additional problems, but some of the cosmetic problems would probably be irrelevant when looking at the system in use, in situ. On the other hand, the in vitro sessions failed to identify two critical problems identified in sitro; both problems had to do with the simulated context of the evaluation. As an example, more of the in sitro participants needed to cancel issued commands, but this was not possible in the tested system. This turned out to be a critical problem in the evaluation because the captain had to apply different means of communication to cancel the command. Such issues never came up in the in vitro condition. But some of these issues from the in vitro condition can probably be explained by the low level of physical fidelity as means for extra communication (e.g. radio communication) that were not present during the in vitro tests.

The general validity of the results of our study is limited in a number of ways. First, the number of test participants used in each study implies that we can primarily explore qualitative issues of changing the condition for usability evaluations. We hope that our study can set out avenues for further research. Also, general competences of the test participants varied between the setups, which could have influenced some of the results. This was especially true for the IT skills of some of the participants, as they had never used a handheld device before. Although there may have been high variability within the groups of participants, we tried to mini-

mize variability in an attempt to decrease the effects on our study. Thus, in all groups we had participants that were not very familiar with mobile and handheld devices. Second, conducting an in situ experiment is controversial in itself, as it can be discussed whether it is possible to observe as closely as we did and still call the condition in situ. In our study, we tried to get as close as possible to a real-use situation for the nurses on morning procedure. Our presence could possibly have influenced some of their behavior and interaction with the mobile devices, and this would have influenced the collected data. This influence is difficult to avoid in the adapted setup, but ethnographic studies involving interviews and observations could address this issue even further. Third, even though our cases are rather different, additional cases could verify the applicability of the in sitro condition. Again, we hope that this would inspire additional study of practical applicability of the condition.

With the increased levels of complexity of mobile technologies and use situations for these mobile technologies, we will probably need additional and innovative ways of evaluating such technologies in the future. In addition, practitioners will eventually start to request methods, heuristics, and guidelines for testing the usability of these mobile technologies. Our research in this article is an attempt to add to this body of knowledge. One possible avenue for further research on this topic could be the determination of what kind of fidelity is needed when evaluating different kinds of mobile technologies. Our study showed that high fidelity in controlled experiments (in sitro) increased the number and types of identified usability problems.

REFERENCES

Abowd, G., & Mynatt, E. (2000). Charting past, present and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction, 7,* 29–58.

Alexander, A. L., Brunyé, T., Sidman, J., & Weil, S. A. (2005, November). *From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in PC-based simulations and games*. DARWARS Training Impact Group, Defense Advanced Research Projects Agency, Arlington, VA. Available from http://www.darwars.net/press/research.html

Allen, J. A., Hays, R. T., & Buffordi, L. C. (1986). Maintenance training, simulator fidelity, and individual differences in transfer of training. *Human Factors, 28,* 497–509.

Amaldi, P., Satinder G. P., Fields, B., & Wong, W. (2005). In-use, in-situ: Extending Field Research Methods Workshop. Interaction Design Centre, Middlesex University, and the BCS HCI Education & Practice SubGroup, BCS HCI Group. Retrieved December 8, 2006, http://www.cs.mdx.ac.uk/research/idc/news_archive/in_use.html.

Andersen, P. B. (2000). *Communication and work on maritime bridges* (CHMI Research Rep. No. CHMI-1-2000). Denmark: Center for Human–Machine Interaction, Århus University. Retrieved from http://www.cs.auc.dk/~pba/ElasticSystems

Arshad, U., Mascolo, C., & Mellor, M. (2003). *Exploiting mobile computing in health-care.* Demo session of the 3rd International Workshop on Smart Appliances, ICDCS03.

Bailey, R. W., Allan, R. W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, HFES*, 409–413.

Barkhuus, L., & Dey, A. (2003). Is context-aware computing taking control away from the user? Three levels of interactivity examined. *Proceedings of the UbiComp2003 conference, LNCS 2864,* 149–156.

Bohnenberger, T., Jameson, A., Krüger, A., & Butz, A. (2002). Location-aware shopping assistance: Evaluation of a decision-theoretic approach. *Proceedings of Mobile HCI 2002*, 155–196.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6,* 188–205.

Chincholle, D., Goldstein, M., Nyberg, M., & Erikson, M. (2002). Lost or found? A usability evaluation of a mobile navigation and location-based service. *Proceedings of Mobile HCI 2002*, 211–224.

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

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259–294.

Driskell, J. E., Johnston, J. H., & Salas, E. (2001). Does stress training generalize to novel settings? *Human Factors, 43,* 99–110.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing.* Exeter, UK: Intellect.

Esbjörnsson, M., Juhlin, O., and Östergren, M. (2003). Motorcyclists using Hocman–Field Trials on mobile interaction. *Proceedings of the 5th International Mobile HCI 2003 conference*, 32–44.

Graham R., & Carter C. (1999). Comparison of speech input and manual control of in-car devices while on-the-move. *Proceedings of the Second Workshop on Human Computer Interaction with Mobile Devices, Mobile HCI 1999.*

Harman, H. H. (1961). *Simulation: A survey* (Rep. No. SP–260). Santa Monica, CA: System Development Corporation.

Hays, R. T., & Singer, M. J. (1989). *Simulation fidelity in training system design*. New York: Springer

Henderson, R., Podd, J., Smith, M., & Varela-Alvarez, H. (1995). An examination of four user-based software evaluation methods. *Interacting with Computers, 7,* 412–432.

Johnson P. (1998). Usability and mobility; Interactions on the move. *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices.*

Kaplan, S. M., & Fitzpatrick, G. (1997). Designing support for remote intensive-care telehealth using the locales framework. *Proceedings of DIS'97, ACM*, 173–184.

Karat, C. M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *Proceedings of CHI'92,* 397–404.

Kjeldskov, J., & Graham, C. (2003). A review of mobile HCI research methods. *Proceedings of the 5th International Mobile HCI 2003 conference*, 317–335.

Kjeldskov, J., & Skov, M. B. (2003). Creating a realistic laboratory setting: A comparative study of three think-aloud usability evaluations of a mobile system. *Proceedings of Interact 2003,* 663–670.

Kjeldskov, J., & Stage, J. (2003). The process of developing a mobile device for communication in a safety-critical domain. In *Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction, Interact 2003. Zurich, Switzerland* (pp. 264–271). IOS Press.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies, 60,* 599–620.

Lai, J., Cheng, K., Green, P., & Tsimhoni, O. (2001). On the road and on the Web? Comprehension of synthetic speech while driving. *Proceedings of CHI'2001,* 206–212.

Luff, P., & Heath, C. (1998). Mobility in collaboration. *Proceedings of CSCW'98,* 305–314.

Mayer, J. D., & Volanth, A. J. (1985). Cognitive involvement in the mood response system. *Motivation & Emotion, 9,* 261–275.

Molich, R. (2000). *Usable Web design*. Ingeniøren | bøger.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., et al. (1998). Comparative evaluation of usability tests. *Proceedings of the Usability Professionals Association Conference,* 189–200.

Nielsen, J. (1993). *Usability engineering.* San Francisco, CA: Morgan Kaufmann.

Nielsen, M. (2000). *Letting go the lines: Departure from the Felixstowe harbour* (CHMI Research Rep. No. CHMI–4–2000). Denmark: Center for Human–Machine Interaction, Århus University. Retrieved from http://www.cs.auc.dk/~pba/ElasticSystems

Pascoe, J., Ryan, N., & Morse, D. (2000). Using while moving: HCI issues in fieldwork environments. *Transactions on Computer-Human Interaction, 7,* 417–437.

Petrie, H., Johnson, V., Furner, S., & Strothotte, T. (1998). Design lifecycles and wearable computers for users with disabilities. *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices.*

Pirhonen, A., Brewster, S. A., & Holguin, C. (2002). Gestural and audio metaphors as a means of control for mobile devices. *Proceedings of CHI'2002*, 291–298.

Pospischil, G., Umlauft, M., & Michlmayr, E. (2002). Designing LoL@, a mobile tourist guide for UMTS. *Proceedings of Mobile HCI 2002,* 140–154.

Rantanen, J., Impio, J., Karinsalo, T., Reho, A., Tasanen, M., & Vanhala, J. (2002). Smart clothing prototype for the Artic environment. *Personal and Ubiquitous Computing, 6,* 3–16.

Roulet, E., Fisch, I., Junier, T., Bucher, P., & Mermod, N. (1998). Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *Silica Biology*, *1*, 21–28.

Rubin, J. (1994). *Handbook of usability testing.* New York: Wiley.

Salvucci, D. D. (2001). Predicting the effects of in-car interfaces on driver behaviour using a cognitive architecture. *Proceedings of CHI'2001,* 120–127.

Sanders, A. F. (1991). Simulation as a tool in the measurement of human performance. *Ergonomics, 34,* 995–1025.

Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *Transactions on Computer-Human Interaction, 7,* 353–383.

Skov, M. B., & Høegh, R. T. (2005). Supporting information access in a hospital ward by a context-aware mobile electronic patient record. *Personal and Ubiquitous Computing*, *10*, 205–214.

Vetere, F., Howard, S., Pedell, S., & Balbo, S. (2003). Walking through mobile use: Novel heuristics and their application. *Proceedings of OzCHI 2003,*

Vreuls, D., & Obermayer, R. W. (1985). Human system performance measurement in training simulations. *Human Factors, 27,* 214–250.

Wingender, E. (1998). ISB: Just another journal? *In Silico Biology. An International Journal on Computational Molecular Biology, 1*.

Zhao, J., Stevens, R. D., Wroe, C. J., Greenwood, M., & Goble, C. A. (2004). The origin and history of in silico experiments. Proceedings of the UK e-Science All Hands Meeting,