ELSEVIER

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Int. J. Human-Computer Studies 60 (2004) 599–620

International Journal of
Human-Computer
Studies

www.elsevier.com/locate/ijhcs

New techniques for usability evaluation of mobile systems

Jesper Kjeldskov, Jan Stage*

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

Received 5 November 2003; accepted 14 November 2003

Abstract

Usability evaluation of systems for mobile computers and devices is an emerging area of research. This paper presents and evaluates six techniques for evaluating the usability of mobile computer systems in laboratory settings. The purpose of these techniques is to facilitate systematic data collection in a controlled environment and support the identification of usability problems that are experienced in mobile use. The proposed techniques involve various aspects of physical motion combined with either needs for navigation in physical space or division of attention. The six techniques are evaluated through two usability experiments where walking in a pedestrian street was used as a reference. Each of the proposed techniques had some similarities to testing in the pedestrian street, but none of them turned out to be completely comparable to that form of field-evaluation. Seating the test subjects at a table supported identification of significantly more usability problems than any of the other proposed techniques. However a large number of the additional problems identified using this technique were categorized as cosmetic. When increasing the amount of physical activity, the test subjects also experienced a significantly increased subjective workload.

1. Introduction

Usability evaluation of systems for stationary computers has grown to be an established discipline within human–computer interaction. Debates are still taking place, but they are often based on a shared understanding of basic concepts. For example, there is a basic distinction between field and laboratory-based evaluations. The majority of literature accepts that both approaches are important and necessary,

*Corresponding author. Tel.: +45-9635-8914; fax: +45-9815-9889.
E-mail addresses: jesper@cs.auc.dk (J. Kjeldskov), jans@cs.auc.dk (J. Stage).

1071-5819/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2003.11.001

and for each of them many authors have contributed with methods and evaluation techniques as well as empirically documented experience with their use.

Extensive guidelines exist that describe how usability evaluations in laboratory settings should be conducted (e.g. Nielsen, 1993; Rubin, 1994; Dumas and Redish, 1999). This is complemented with experimental evaluations of the relative strengths and weaknesses of different techniques that can be applied in a usability evaluation (e.g. Bailey et al., 1992; Karat et al., 1992; Henderson et al., 1995; Molich et al., 1998).

Established concepts, methodologies, and approaches in human–computer interaction are being challenged by the increasing focus on systems for wearable, handheld, and mobile computing devices. This move beyond office, home, and other stationary use settings has created a need for new approaches to design and evaluate useful and usable systems (see e.g. Luff and Heath, 1998).

Mobile systems are typically used in highly dynamic contexts. Moreover, their use often involves several people distributed in the user's physical surroundings (Danesh et al., 2001; Kjeldskov and Skov, 2003a, b). Therefore, field-based evaluations seem like an appealing, or even indispensable, approach for evaluating the usability of a mobile system. Yet evaluating usability in the field is not easy (Nielsen, 1998; Brewster, 2002). Three fundamental difficulties are reported in the literature. Firstly, it can be complicated to establish realistic studies that capture key situations in the use-context described above (Pascoe et al., 2000; Rantanen et al., 2002). Secondly, it is far from trivial to apply established evaluation techniques such as observation and think-aloud when an evaluation is conducted in a field setting (Sawhney and Schmandt, 2000). Thirdly, field evaluations complicate data collection and limits control since users are moving physically in an environment with a number of unknown variables potentially affecting the set-up (Johnson, 1998; Petrie et al., 1998).

In a laboratory setting, these difficulties are significantly reduced. When usability evaluations are conducted in a laboratory setting, experimental control and collection of high quality data is not a problem. Yet one of the drawbacks of this setting is the lack of realism. Existing approaches to laboratory-based usability evaluations of stationary computer systems try to solve this problem by recreating or imitating the real context of use in the laboratory by, for example, furnishing it as an office (Rubin, 1994). However, when mobile systems are evaluated in a laboratory setting, mobility of the user and activities in the user's physical surroundings can be difficult to recreate realistically (Pirhonen et al., 2002; Thomas et al., 2002). Evaluation of mobile system usability is increasingly being reported (see e.g. Brewster and Murray, 2000; Sharples et al., 2002). A recent survey of mobile human–computer interaction research has shown that laboratory experiments are presently the most prevalent method for evaluating mobile systems (Kjeldskov and Graham, 2003). This study reveals that 41% of the surveyed research on mobile human–computer interaction between 2000 and 2002 involves evaluations of system designs. 71% of these evaluations are conducted by means of laboratory experiments and very few of these laboratory evaluations involve special techniques being applied to meet the challenges of evaluating a *mobile* system.

In the light of this, the purpose of this paper is to explore new techniques for evaluating the usability of mobile systems in laboratory settings that addresses the limitations discussed above while preserving the advantages of a laboratory experiment.

In Section 2, we present the results from a comprehensive study of existing literature on usability testing of mobile systems. Section 3 presents ideas for new techniques that can recreate or imitate real-world use situations of a mobile system in a laboratory setting. Section 4 describes the design of two experiments, inquiring into the qualities of these techniques. The purpose of these experiments was to explore and compare several techniques, rather than evaluating a single technique. Sections 5 and 6 present and discuss the results from these experiments and Section 7 provides the conclusion.

2. Related work

The literature on human–computer interaction contains a number of contributions on techniques for evaluating the usability of mobile systems. In order to identify proposals for new techniques, we have searched part of that literature.

2.1. Literature survey

To identify literature that deal with usability evaluation of mobile systems, we conducted a systematic literature search in the following journals and conference proceedings.

- Proceedings of the International Conference on Mobile HCI: 1998, 1999, 2001.
- Proceedings of the ACM Conference on Computer-Human Interaction: 1996–2002.
- ACM Transactions on Computer-Human Interaction (TOCHI): 1996–2002.

While books and other journals and conference proceeding series exists, which also report interesting research in mobile human–computer interaction (see e.g. Kjeldskov and Graham, 2003), we found that this selection provided a representative overview of state-of-the art in usability evaluation techniques for mobile systems. A total of 636 papers were examined, resulting in the identification of 114 papers dealing with human–computer interaction for mobile systems. These papers are categorized in Table 1.

The 2 papers in category A deal with general aspects of usability evaluations of mobile systems and provide practical advice. In the 11 papers in category B, usability evaluations were carried out based on simulations of mobile systems on desktop personal computers. The 44 papers in category C typically deals with design of mobile applications, and employs traditional usability evaluation methods such as heuristic inspection and think-aloud in laboratory settings. Many of them employ a technique where test subjects are being seated at a table in order to test a device that

Table 1
Distribution of papers dealing with human–computer interaction for mobile systems

		ACM CHI	ACM TOCHI	Mobile HCI	Total
A	General aspects of usability evaluations	0	1	1	2
B	Usability evaluations on device simulator	5	2	4	11
C	Usability evaluations with traditional techniques	34	3	7	44
D	Usability evaluations with new techniques	3	0	3	6
E	Usability evaluations not described	6	0	9	15
F	No usability evaluations performed	3	3	30	36
	Total	51	9	54	114

was intended for use in a mobile situation. Contrary to this, the 6 papers in category D present and apply new techniques for usability evaluations in order to reflect or recreate a mobile use situation. Below, we describe these proposals for new techniques in more detail.

The 15 papers in category E mention that usability evaluations have been carried out but do not describe them. In the 36 papers in category F, no usability evaluations were performed.

2.2. Proposed techniques

The six papers in category D employ new and different techniques for increasing the realism of the evaluation situation. In these papers, there are two basic categories of techniques.

In the first category, the test subjects were required to walk while using the mobile system being evaluated. This would either take place on a treadmill or on a specifically defined track in a laboratory set-up (Pirhonen et al., 2002) or on a real world route, which also involved way finding etc. (Petrie et al., 1998). Both of these settings facilitated the collection of a magnitude of qualitative and quantitative data such as task completion time, error rate, heart rate, perceived cognitive workload and deviation from preferred walking speed, etc.

In the second category, the test subjects were using a mobile system while driving a car simulator. The type of car simulator that was used ranges from low-fidelity personal computer-based simulations (Graham and Carter, 1999; Koppinen, 1999) to high-fidelity simulators with large projection screens involving real dashboards (Lai et al., 2001) or even real cars (Salvucci, 2001). This technique does not involve the user being physically mobile to the same degree as when walking, but it facilitates the evaluation of mobile system use while simultaneously engaged in a demanding cognitive activity. The simulator-based technique facilitated both quantitative and qualitative data to be collected and a huge number of test sessions to be conducted within limited time frames.

Only two of the six papers in category D (Graham and Carter, 1999; Pirhonen et al., 2002) employ multiple techniques or variations of proposed techniques. This is done to systematically measure their relative applicability and ability to support the identification of usability problems in the mobile systems being evaluated.

Overall, the literature study reveals that only a limited amount of human–computer interaction research involves usability evaluations of mobile systems. Out of 114 papers dealing with human–computer interaction for mobile systems, less than half of them report results from a usability evaluation of the presented design (category C and D combined). This is consistent with the tendency identified in (Kjeldskov and Graham, 2003). Furthermore, the majority of usability evaluations of mobile systems employ only traditional techniques (category C) and little variety exists within the studies employing new techniques. This raises two key questions about usability evaluations of mobile systems. (1) Are traditional techniques optimal for evaluating the usability of mobile systems? (2) What new techniques might be suggested?

3. Ideas for new techniques

To complement the traditional techniques for usability evaluation, we have developed a number of alternatives. In order to make this effort more systematic, we used the literature on mobility as our point of departure. The problem is, however, that much of this literature deal with mobile systems on a level that is much more abstract than the physical activity of a person using a mobile system in a specific context. For example, mobility has been described in terms of application, mobility type, and context, where mobility type then is sub-divided into visiting, travelling, and wandering (Kristoffersen and Ljungberg, 1999). Yet when looking at the different ways in which a mobile phone is used in order to recreate these situations in a laboratory, such a framework is not particularly helpful, and it is very difficult to use it for generating alternatives to traditional usability evaluation techniques.

Based on these experiences, we changed our focus to theories on human information processing and their description of attention and conscious action (a summary can be found in Preece et al., 1994). Based on these theories, we developed two different frameworks for mobile use.

3.1. Framework A

Framework A focused on the different ways in which a user could be moving physically while using a mobile system. The following two dimensions describing this are as follows.

- *Type of body motion*: none, constant, varying.
- *Attention needed to navigate*: none, conscious.

By juxtaposing these two dimensions we ended up with a two by three matrix of different overall configurations for incorporating mobility in laboratory usability

Table 2
The five configurations based on motion and need for navigation

Body motion	Attention needed to navigate	
	None	Conscious
None	1. Sitting at a table or standing	n/a
Constant	2. Walking on a treadmill with constant speed or stepping on a stepping machine	4. Walking at constant speed on a track that is changing because obstructions are moved
Varying	3. Walking on a treadmill with varying speed	5. Walking at varying speed on a track that is changing because obstructions are moved

evaluation set-ups (Table 2). If there is no motion, no navigation in physical space is necessary, which leaves one cell empty.

In relation to the previous research presented in Section 2, configuration 1 in Table 2 is the traditional evaluation situation, where a user is sitting at a table or standing still while using a mobile system. This corresponds to the research in category C in Table 1. The user is not moving through physical space, but the configuration is still being used in many usability evaluations of mobile systems. The research in category D in Table 1 can also be related to the configurations in Table 2. For example, it has been suggested to use a hallway with fixed obstructions as the experimental setting, which corresponds to configuration 4, or a stepping machine that allows the use to walk without moving (Pirhonen et al., 2002), which corresponds to configuration 2 in Table 2 above.

3.2. Framework B

Framework B was based on the notion of divided attention. When people are using a mobile system while being mobile, their attention is divided between physical motion and the use of the system. Similar to the studies applying car simulators for usability evaluations of mobile systems, we thus aimed at creating a configuration that replicated a division of attention between a demanding cognitive task and the use of a mobile system. Unlike studies determining, for example, the deviation from the user's preferred walking speed during the evaluation (Pirhonen et al., 2002) we were not interested in measuring the user's performance on the secondary task. Our aim of the dual task approach was only that it should serve as a distracting factor in the laboratory experiment in order to simulate divided attention directing the user's focus away from the use of the mobile system in a real world setting.

4. Experimental design

We conducted two different experiments, each based on one of the two frameworks described above. The two experiments evaluated two different types of mobile systems for text-based communication. In this section, we describe the design of these experiments.

4.1. Experiment A

The first experiment was based on framework A and the five configurations described in Table 2 (Beck et al., 2002). The purpose of this experiment was to inquire into the relative strengths and weaknesses of the different configurations when used as techniques for usability evaluations in a laboratory. In addition, we wanted to compare these to a typical use situation in the field. Thus the experiment involved the following six techniques, of which the first five match the five configurations described earlier.

1. Sitting on a chair at a table.
2. Walking on a treadmill at constant speed.
3. Walking on a treadmill at varying speed.
4. Walking at constant speed on a course that is constantly changing.
5. Walking at varying speed on a course that is constantly changing.
6. Walking in a pedestrian street. This embodies a typical use situation and is intended to serve as a reference for the other techniques.

We conducted a series of usability evaluations employing each of the six techniques above. In each evaluation, the user solved a number of specified tasks using a mobile system. The first five techniques were used in a usability laboratory (Fig. 1). The sixth technique was used in the field. For configurations 4 and 5, the user had to walk a sequence of three different courses (depicted in Fig. 2).

The mobile system used for the experiment was an experimental short messaging service (SMS) application for the Compaq iPAQ personal digital assistant. This application provided the user with facilities for sending and receiving short text messages. In addition, the application complied with the specification of enhanced message service (EMS) (Ericsson, 2001), which enables exchange of small sound clips and pictures as part of a message. The application was specially designed for the experiment so that user actions and performance could be accurately measured through a dedicated monitoring application running in the background. We decided to use the short message service because it is widely used among mobile users and because it is highly interactive, involving both reading on the screen and typing in letters on a keyboard.

Each test subject was presented with five tasks involving sending and receiving short messages. While solving the tasks, the test subjects were required to think-aloud. In order to keep the time schedule, we decided to allocate ten minutes to each test subject. When 10 min had passed, the evaluation was stopped even if the test subject had not completed all tasks. For each evaluation, we collected three types of data.

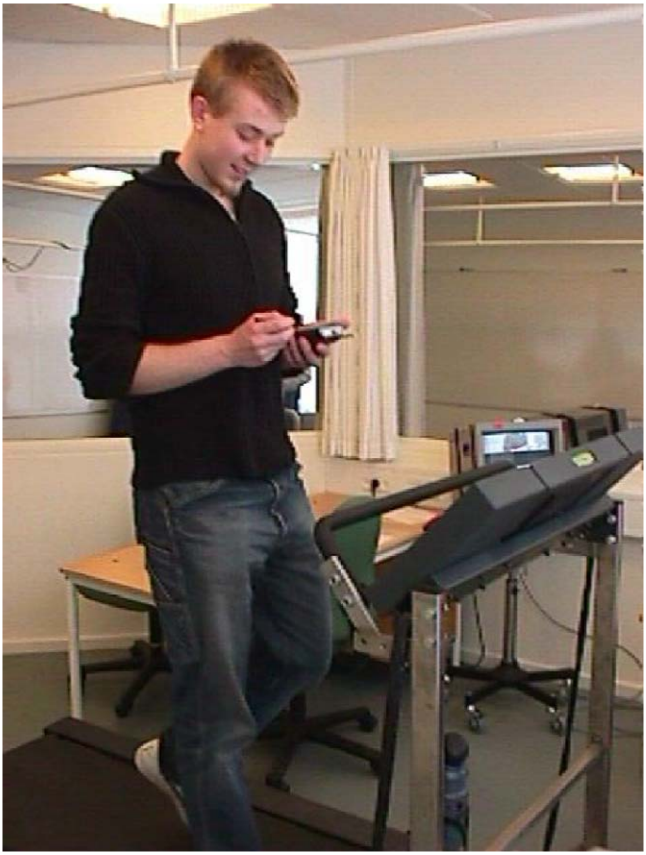

Fig. 1. Test subject walking on a treadmill in the usability laboratory.

- *Usability problems*: all evaluations in both the laboratory and the field were recorded on video. After the evaluation, the video recordings were analysed to produce a list of usability problems.
- *Performance*: a dedicated monitoring application automatically collected data about user interaction and time spent on each task.
- *Workload*: immediately after each evaluation, a NASA task load index (TLX) test was conducted with the test subject. This test assesses the user's subjective experience of the overall workload and the factors that contribute to it (Hart and Staveland, 1988; NASA).

The overall hypothesis was that, the ideal laboratory techniques would not differ from walking in a pedestrian street (technique 6) in terms of these three measures.

A test monitor managed each individual evaluation. Three experienced usability experts served as evaluators. In order to ensure that they were conducted

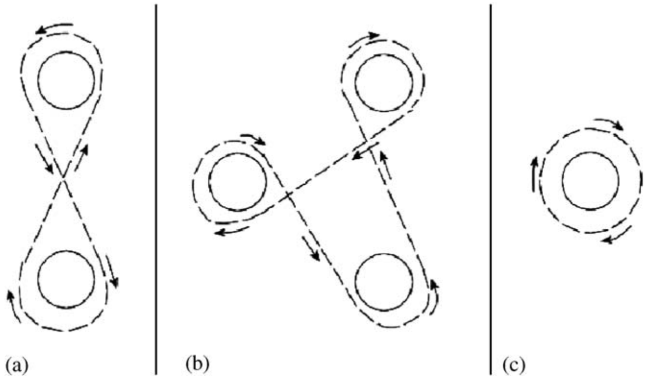

Fig. 2. The sequence of the different courses to be walked in the laboratory (a–c).

consistently, the evaluators remained the same throughout all evaluations conducted using a specific technique.

We were not able to carry out an unlimited number of evaluation sessions. Therefore, the number of test subjects used for each technique had to be a trade-off between the number of evaluations with each technique and the number of techniques we would be able to evaluate at all. Our aim was to explore a number of different techniques and not only evaluate one or two (cf. Section 1). For that reason we had to minimize the number of test subjects used for evaluating by means of each technique. Key literature on usability evaluations suggests that it is possible to find around 80–85% of all usability problems if five test subjects are used (Virzi, 1992; Nielsen, 2000). Also, it has been argued that four to five test subjects are sufficient to gain an overall idea about the usability, but to avoid missing a critical problem, at least eight subjects should be used (Rubin, 1994). Based on this, we planned to use eight test subjects for each technique, amounting to a total of 48 test subjects.

The test subjects were all students of Informatics or Computer Science at the University of Aalborg, Denmark. They were demographically homogeneous, realistic users of the application, and easy to contact. We contacted students who were at the end of their first or second year on these two educations. They answered a questionnaire about personal characteristics, experiences with mobile phones and personal digital assistants, their knowledge about the short messaging service, and their prior involvement in usability evaluations. Based on their responses we distributed them on the different techniques with the intention of avoiding bias. Due to test subjects cancelling or not showing up, and our aim to conduct the same number of tests with each technique, we ended up conducting six tests with each of the six techniques, thus involving a total of 36 test subjects.

After the evaluation sessions, the three usability evaluators analysed the video recordings individually in random order and produced three lists of usability problems with severity ratings of critical, serious or cosmetic in accordance to the definition proposed by Molich (2000). Following this, the evaluators met and discussed their individual problem lists until consensus on one complete list was reached.

4.2. Experiment B

To complement experiment A, we designed and conducted a different experiment based on framework B (Jacobsen et al., 2002). The purpose of this experiment was to compare the extent to which laboratory and field evaluations supported the comparison of two different mobile phones. Experiment B involved two different techniques.

1. Using a mobile system while playing a computer game requiring the player to move around physically on a mat placed on the floor.
2. Using a mobile system while walking in a pedestrian street (a typical use situation) serving as a reference for the other technique.

The computer game used in the first technique was the "Jungle Book Groove Party" for Sony PlayStation 2. When playing this game, the user steps on different active areas on a "dance mat" according to sequences shown on a monitor. In all evaluations, the test subject played the game on the easiest level. The idea was to force the user into a situation with clearly divided attention between the use of the mobile system and playing the game. The second technique was similar to the pedestrian street technique mentioned in experiment A (Fig. 3).

We conducted a series of usability evaluations employing each of the two techniques above to evaluate the usability of two different mobile phones. In each evaluation, the user solved a number of tasks using one of the two mobile phones.

The two mobile phones were the Nokia 3310 and the Nokia 5510 (Fig. 4). They have comparable functionalities, but the keyboards are very different. The Nokia 3310 has a typical mobile phone keyboard with one digit and three or more letters assigned to each key. The Nokia 5510 has a full keyboard with only one character

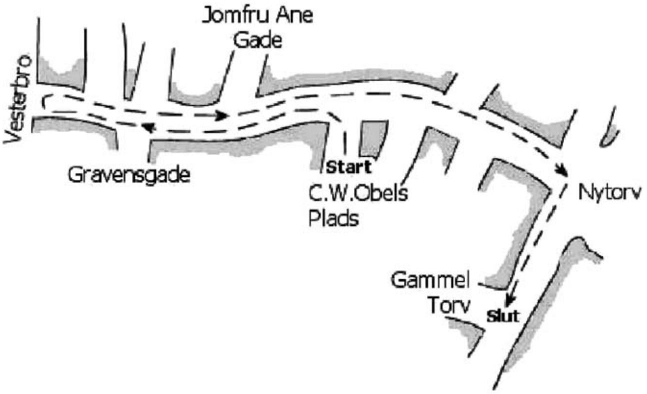

Fig. 3. The route to be walked in the evaluations in the pedestrian street.

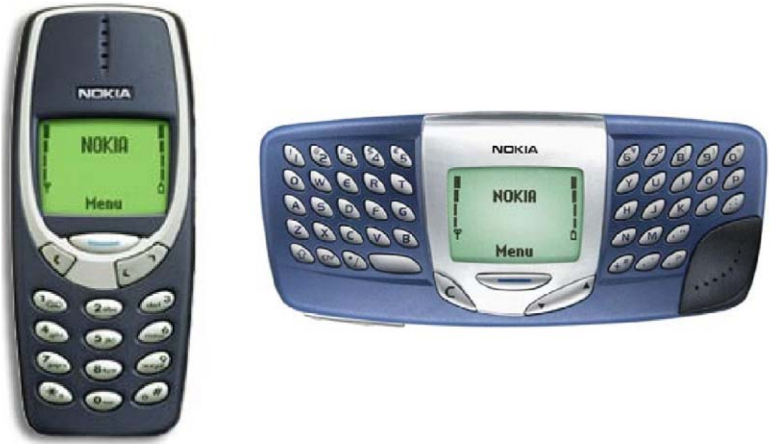

Fig. 4. The Nokia 3310 and Nokia 5510.

assigned to each key. Again, we decided to focus on the use of the short messaging service because this is a widely used service among mobile users and because it requires extensive text input as well as reading on the screen. In each evaluation we collected two types of data.

- *Usability problems*: the evaluations in the laboratory were recorded on video. In the pedestrian street evaluations only audio was recorded and an observer took notes. After the evaluations, the video and audio recordings and notes were analysed to produce a list of usability problems.
- *Performance*: from the video and audio recordings, the time spent on solving each task was measured.

The overall hypothesis of the experiment was that the two techniques would come out with similar results in terms of the data collected.

Four persons took turns in serving as evaluators. The test subjects were a mixture of grammar school students, university students, and business employees. As we only had 12 test subjects available for experiment B, we decided to let them use both mobile phones in counterbalanced order. The test subjects were divided into two groups according to their frequency of mobile phone and short messaging service use. When the evaluations were carried out, one test subject did not show up, and two test subjects were only able to participate in one test.

The test subjects were presented with two tasks; a trial task and an evaluation task. The trial task was completed while standing still, without walking or using the dance mat. The purpose of the trial task was to make sure that all subjects had used the specific type of mobile phone being evaluated at least once. The evaluation task involved the same functionality as the trial task, but the sequence and data were different.

After the evaluations, the four usability evaluators analysed the video recordings individually in random order and produced four lists of usability problems with severity ratings of critical, serious or cosmetic (Molich, 2000). The individual problem lists were then merged into one complete list through discussions among the four evaluators until consensus was reached.

5. Results

This section presents the key results from the two experiments described above.

5.1. Experiment A

In experiment A, we collected data about identification of usability problems, performance and workload.

5.1.1. Usability problems

The primary basis for evaluating a technique should be the number of usability problems it helps identifying. Thus, we have analysed the collected data in order to evaluate the extent to which each technique supports identification of usability problems.

Our hypothesis was that the best laboratory technique would identify a number of usability problems similar to the number identified in the pedestrian street condition (technique 6). Table 3 is based on the number of usability problems identified with the 36 test subjects. It shows the mean number of identified usability problems along with the standard deviation, distributed on technique. This is also illustrated in Fig. 5 below.

An analysis of variance of these numbers shows that the difference between the means are highly significant ($F_{5,30} = 2.53$, $p = 0.001$) and the use of Fisher's least significant difference test supports the conclusion that sitting at a table (technique 1) differs from the rest of the techniques. This means that the test subjects in the sitting condition supported identification of significantly more usability problems than with any of the other techniques.

Table 4 shows the number of usability problems categorized from severity ratings and distributed on techniques. For severity rating, we used the criteria proposed by

Table 3
Mean numbers and standard deviations of usability problems identified by each of the six techniques

	Technique					
	1	2	3	4	5	6
Mean	10.8	7.5	6.7	6.7	5.2	6.3
Std. deviation	1.6	1.5	2.0	2.4	2.4	2.1

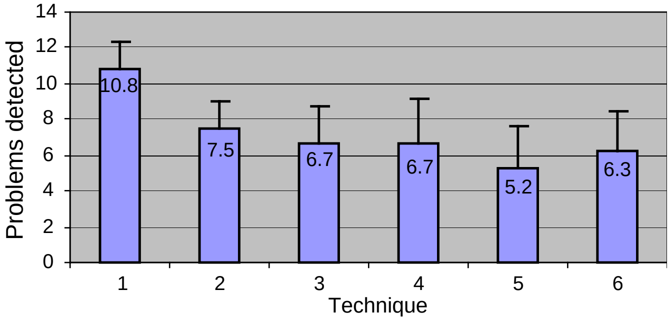

Fig. 5. Number of usability problems detected with each of the six techniques.

Table 4
Number of identified usability problems

	Technique						Combined
	1	2	3	4	5	6	
Critical	4	4	3	4	3	3	4
Serious	11	11	9	9	9	8	17
Cosmetic	19	8	8	8	6	12	32
Sum	34	23	20	21	18	23	53

Molich (2000). In total, 53 unique usability problems were identified across all techniques, consisting of 32 cosmetic problems, 17 serious problems and 4 critical problems.

No single technique supported the identification of all usability problems. In the sitting condition (technique 1), a total of 34 problems were identified. All of the others supported the identification of roughly half the usability problems. Looking at critical and serious problems, the six techniques supported the identification of nearly the same number. The main difference between the sitting condition and the other techniques relates to the number of cosmetic problems. Thus more than double the number of cosmetic problems was identified in the sitting condition than when using any of the other laboratory techniques.

5.1.2. *Performance*

We expected clear differences between the performances that test subjects would achieve with the different evaluation techniques. The main measure was the time spent on solving each of the five tasks. Our hypothesis was that the users who employed the best techniques would have similar performance to those walking in the pedestrian street (technique 6).

An analysis of variance of the time spent on each task with each technique did not enable us to identify any systematic differences between the techniques. The technique with the fastest task completion time changed from task to task. We also

Table 5
Subjective experience of workload with the different techniques

	Technique					
	1	2	3	4	5	6
Mental demands	29	75	204	126	185	148
Physical demands	92	117	112	118	127	194
Effort	52	163	106	228	178	186
Overall workload	27	35	48	55	48	54

analysed the number of completed tasks, the number of wrong and undetected pressings of buttons, and the number of requests from the subject to have the task description repeated without finding any differences that could be clearly attributed to the different techniques.

5.1.3. Workload

The data on workload exhibited more difference between the techniques. Based on the theories behind the new techniques our hypothesis was that the workload would differ significantly as more body motion or attention was needed.

Table 5 shows the calculated workload numbers for the overall workload and three of the six contributing factors. The last three contributing factors; timing demands, own performance, and frustration did not provide any significant results.

Mental demands are described in the NASA task load index (TLX) as the amount of mental activity, e.g. thinking or use of memory, required to perform a piece of work. We were expecting to see an increase in these demands, as more attention was required in the evaluations. As can be seen in Table 5, the numbers do seem very different, ranging from 29 to 204 and increasing as the techniques become more complicated. However, a variance analysis did not show any significant difference between the techniques ($F_{5,30} = 1.91$, $p = 0.12$). The reason for this is that the test subjects within each technique rated this factor very differently—for one technique the rating ranged from 15 to 85.

A pairwise comparison of technique 1–6 does show some significant differences though: compared to all but walking on a treadmill at constant speed (technique 2), the sitting condition (technique 1) demands significantly less mental activity, which seems to confirm that the new laboratory techniques and the pedestrian street evaluations require more attention from the test subjects. However further comparisons between the techniques in relation to physical demands, effort and overall workload do not reveal any significant results, aside from walking on a treadmill at constant speed (technique 2) compared to walking on a treadmill at varying speed (technique 3).

Physical demands mean how much physical activity was required, including both walking, dragging an icon or pushing a button. We expected to see an increase in this for the techniques, which involved more body motion. However, as before, a variance analysis shows no significant differences ($F_{5,30} = 0.48$, $p = 0.79$). Despite the seemingly big difference between e.g. walking on a treadmill at constant speed (technique 2) and walking in a pedestrian street (technique 6) a pairwise comparison of the techniques show no significant differences either.

Effort is explained as a combination of the mental and physical demands and for this our variance analysis shows significant difference between the techniques ($F_{5,30} = 3.27, p = 0.02$). Fisher's least significant difference test identifies that sitting requires significantly less effort than the other techniques. When we compare the figures for techniques with the same sort of attention but with constant versus varying speed, i.e. walking on a treadmill (technique 2 versus 3) or walking on a changing track (technique 4 versus 5), the test subjects seem to feel that constant speed requires more effort. However, the difference is not significant. The reason may be that the varying speed included some intervals with low speed where the test subjects got a chance to relax, before the speed increased again.

The overall workload exhibits a very significant difference between the techniques ($F_{5,30} = 4.14, p < 0.01$). Fisher's least significant difference test reveals no significant difference between sitting (technique 1) and walking on a treadmill at constant speed (technique 2), which indicates that the motion at constant speed is not experienced very different from sitting. The reason may be that walking at constant speed quickly becomes automatic and therefore not requiring much more attention than sitting at a table. On the other hand, walking at varied speed (technique 3) is significantly different from sitting. Walking on a changing course (techniques 4 and 5) and in a pedestrian street (technique 6) are also significantly different from sitting and walking at constant speed, which implies that varied speed and conscious attention do in fact put a greater workload on the test subject. However there is no significant difference between the two techniques involving running or between the two techniques where the subjects are walking on a changing track.

It is impossible to pick a single technique that fully resembles the workload for the pedestrian street technique. For some factors having a need for both varying body motion and conscious attention seem to simulate the pedestrian street the best while in others, the less complicated techniques seem to be better.

5.2. Experiment B

The usability problems found from experiment B are shown in Table 6. Again, severity ratings are based on the criteria proposed by Molich (2000).

Table 6
Number of usability problems detected with each technique

	Dance mat				Pedestrian street			
	Critical	Serious	Cosmetic	Total	Critical	Serious	Cosmetic	Total
Nokia 3310	0	8	3	11	2	4	3	9
Nokia 5510	0	10	11	21	4	10	9	23

Table 7
Average time spent solving the task with the mobile phones (min:s)

	Dance mat			Pedestrian street		
	Subject	Trial	Evaluation	Subject	Trial	Evaluation
Nokia 3310	S1	5:55	3:59	S7	3:10	3:38
	S2	2:48	3:52	S8	5:27	4:12
	S3	3:10	3:19	S9	4:20	2:44
	S4	6:52	7:44	S10	3:04	3:30
	S5	3:28	4:34	S11	3:05	2:40
	S6	—	—	S12	—	—
	Mean	4:27	4:42	Mean	3:49	3:21
Nokia 5510	S1	6:43	5:41	S7	6:50	3:57
	S2	4:27	4:30	S8	3:57	3:15
	S3	2:29	2:35	S9	3:49	2:44
	S4	7:09	7:02	S10	—	—
	S5	8:40	6:37	S11	5:49	4:01
	S6	7:52	5:46	S12	—	—
	Mean	6:13	5:22	Mean	5:06	3:29

For both mobile phones, a comparable number of usability problems are identified across techniques (11 and 9 for the Nokia 3310, and 21 and 23 for the Nokia 5510). However, with the pedestrian street technique the evaluation identified some critical problems, which were not identified in the dance mat condition.

The measures of performance for this experiment are provided in Table 7. Each number is the time taken for a test subject to complete the task in the given context.

This table illustrates that the test subjects walking in the pedestrian street solved the tasks faster than the test subjects that used the dance mat. In the pedestrian street, the evaluation task was completed faster than the trial task. This improvement was expected as the subjects became familiar with the mobile phone. With the dance mat, we have the opposite result. The trial task was solved faster than the evaluation task. This indicates that the dance mat demands more attention from the user than walking in the pedestrian street.

6. Discussion

This section discusses the experiments and issues from them that go beyond the results presented above.

6.1. The sitting technique

An interesting and surprising result of our experiments is that technique 1 (sitting at a table) supports the identification of more usability problems than any other techniques. In this sense, the traditional usability evaluation technique seems superior.

The data on workload indicates a potential reason of this result. Pairwise comparisons of the workload data across the techniques show that sitting (technique 1) compared to all other techniques, except walking on a treadmill at constant speed (technique 2), demands significantly less mental activity.

We have analysed the video recordings from experiment A for consequences of this reduced experience of mental demands. Generally, identification of usability problems is based on the test subjects thinking aloud. If the test subjects talk less, we may miss usability problems. Our video recordings indicate that the test subjects in the sitting condition (technique 1) spent more time and energy thinking aloud and commenting on what they observed compared to the test subjects in the other tests. The test subjects who were sitting down at the table also had energy to comment on all the small things they observed. The test subjects in the evaluations based on the five other techniques were mostly thinking aloud when they observed a larger usability problem.

Theory on human information processing can be used to explain this. Thinking aloud is a conscious action, which requires some amount of attention, much in the same way as motion and navigation. With sitting (technique 1) the users only needed to do one action, which was to solve the tasks. Therefore, these test subjects only had to divide their attention and effort between two actions: solving the task and thinking aloud. With the other five techniques, the users needed to solve the tasks, move physically and navigate. Therefore, these test subjects had to divide their attention and effort between three or more actions; solving the task, moving, navigating, and thinking aloud.

The number of usability problems found in the evaluations provides an indication of the consequences of this different demand. The results show that the techniques involving multiple actions support identification of less usability problems than the techniques with fewer actions. But when the usability problems are divided into the three categories of critical, serious and cosmetic problems, the results show that the major differences between the techniques reside in the amount of cosmetic problems found. This supports our assumption that the test subjects in the sitting condition (technique 1) point out every problem they find as opposed to the test subjects in the other five techniques, who only point out the serious and critical problems they encounter.

6.2. Usability problems and mobility

A detailed analysis of the performance results revealed that the test subjects in techniques with much motion and navigation are more likely to miss a button on the interface. This can happen if a user presses a button but moves the stylus out of the button before releasing it, or if a user unintentionally hits the wrong button.

In experiment A, the test subjects that were sitting at a table (technique 1) missed a button on average 2 times throughout the whole evaluation. In the techniques with motion but no navigation (technique 2 and 4) a button was missed about 3 times per test subject. In the techniques that involved both motion and navigation (techniques 3, 5 and 6), a button was missed on average between 3.5 and 6 times per test subject. This difference between the six techniques is less significant ($F_{5,30} = 2.30$, $p = 0.10$), but it indicates that the techniques involving movement and navigation are better at finding problems concerning the interface layout and the sizes and placement of the individual interface elements.

Problems using the devices and programs can also be found in experiment B. In the pedestrian street evaluation, a total of 14 errors were made (divided on six of the 12 tests) when writing the short messages, whereas only five errors were made in the dance mat test (divided on only three of the 12 tests). Unfortunately, the data collected in this experiment do not allow further enquiry into the causes of this difference as only audio was recorded.

6.3. A changing track

One of the techniques involved walking on a track that was changing (technique 5). The idea of changing the track was to increase the need for attention. However, the pairwise comparisons between mental demands for each technique did not reveal such an effect.

One possible reason for this was discovered during the usability evaluations involving technique 4 and 5. We learned that most of the test subjects just followed the person ahead of them by keeping track at them out of the corner of their eyes. Rather than navigating between the obstructions the test subjects simply followed the person who set the speed and counted on him to avoid walking into anything, thereby reducing the attention needed. Thus the navigation did not appear to be as conscious as we wanted it to be.

This problem may be solved by e.g. making the laboratory set-up even more dynamic with more persons and moving objects.

6.4. Data collection in the field

We designed the pedestrian technique to include systematic data collection. In experiment A, all tests in the pedestrian street were video recorded using a video camcorder as shown in Fig. 6. In experiment B, we only recorded audio but also took written notes.

Collecting high-quality video data in the field turned out to be very difficult. It was not easy to record images of the screen of the iPAQ while walking. In addition, the users often moved their hands in a manner that covered the screen. Furthermore, it was difficult to experience “realistic” pedestrian motion, since the other pedestrians tended to move away from the three persons walking along the street; the test subject, the evaluator and the person operating the video camcorder (see Fig. 6). This problem may be solved by e.g. changing the role of the evaluator and mounting small cameras and microphones on the test subject.

Recording only audio in the field resulted in a lack of detailed data about the interaction with the system. Consequently, we had to rely heavily on written notes

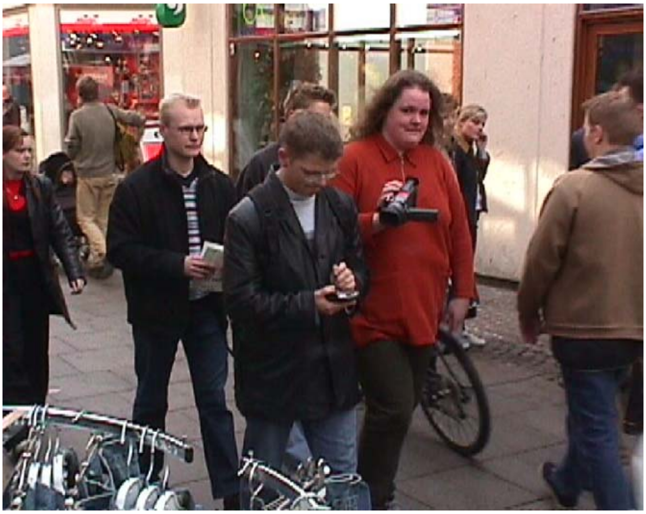

Fig. 6. Usability evaluation in the pedestrian street.

during analysis. Thus while this approach was less obtrusive and easier to carry out in practice, it did not overall prove more valuable than recording the field evaluations on video.

6.5. Involving social context

When the places and environments where mobile systems are being used are compared to the theories that we have used for creating our testing techniques, there is a gap in the experiments. One of these gaps is the integration of social context. For instance, a user of a mobile system may be working with some colleagues while interacting with the mobile system or the user may be working with colleagues through the mobile system. We have only used theories that enable us to see mobility as something that involves motion and navigation. None of the theories cover the social context. Therefore, this aspect has not been a part of our experiments. Future research should investigate further into this matter.

7. Conclusion

This paper has presented six techniques for evaluating the usability of a mobile system in a laboratory setting. The aim was to explore techniques that could facilitate evaluating mobile systems in a controlled environment while being as similar to a real use situation as possible.

Five techniques were developed from a framework that described mobility in terms of physical motion and the amount of attention needed to navigate while moving. A sixth technique was developed to divide the user's attention between conscious actions and the use of the mobile system.

The proposed techniques were evaluated through two experiments. In both experiments, walking in a pedestrian street while using the mobile system being evaluated was used as reference. There were no significant differences between the techniques in terms of user performance. On workload the techniques exhibited significant differences in terms of perceived effort and overall workload. However, there was no single technique that resulted in exactly the same workload as walking in the pedestrian street.

There was only one significant difference in terms of support to identification of usability problems. Sitting at a table, which was the simplest of the six new techniques, was clearly better than any other technique when focusing on identification of usability problems. However, the difference mainly related to cosmetic problems.

Both of the experiments have clear limitations. Each technique has been evaluated with six test subjects, except for three cases in the second experiment, where only four or five test subjects were used. More test subjects would have been desirable. Our aim was to facilitate comparison of several techniques with a limited number of test subjects. A follow-up experiment on selected techniques should increase the number of test subjects.

The proposed framework and subsequent experiments have focused on exploring techniques for recreating challenges of moving physically while interacting with a mobile computer system. Moving physically is, however, only one of many new factors involved in the use of mobile computer systems as opposed to traditional desktop applications. Other relevant factors are the social, physical and temporal context of mobile system use. Some studies exists, which explore how these factors influence usability evaluations and how they can be incorporated into laboratory based evaluation techniques (Lai et al., 2001; Salvucci, 2001; Kjeldskov and Skov, 2003b) but further research and experiments are needed to develop new and refine existing ideas and techniques.

Acknowledgements

The authors thank all test subjects and Elena T. Beck, Morten K. Christiansen, Nikolaj Kolbe, Charlotte S. Jacobsen, Pernille Langvad, Janne J. Sørensen and Heidi B. Thomsen who participated in the conduction of the two experiments.

References

Bailey, R.W., Allan, R.W., Riello, P. 1992. Usability testing vs. heuristic evaluation: a head-to-head comparison. Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, HFES, pp. 409–413.

Beck, E.T., Christiansen, M.K., Kolbe N., 2002. Metoder til Brugbarhedstest af Mobile Apparater. Department of Computer Science, Aalborg University (in Danish).

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

Brewster, S., Murray, R., 2000. Presenting dynamic information on mobile computers. Personal and Ubiquitous Computing 4, 209–212.

Danesh, A., Inkpen, K., Lau, F., Shu, K., Booth, K., 2001. Geney: Designing a Collaborative Activity for the Palm Handheld Computer. Proceedings of CHI'2001, ACM, New York, pp. 388–395.

Dumas, J.S., Redish, J.C., 1999. A Practical Guide to Usability Testing. Intellect, Exeter.

Ericsson Mobile Communications, 2001. Enhanced Messaging Service, white paper, http://www.ericsson.com.au/about/media_center/white_papers/articles/EMS.pdf.

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, Edinburgh, Scotland.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. Elsevier Science Publishers, Amsterdam, pp. 139–183.

Henderson, R., Podd, J., Smith, M., Varela-Alvarez, H., 1995. An examination of four user-based software evaluation methods. Interacting with Computers 7 (4), 412–432.

Jacobsen, C.S., Langvad, P., Sørensen J.J., Thomsen H.B., 2002. Udvikling og Vurdering af Brugbarhedstest til Mobile Apparater. Department of Computer Science, Aalborg University (in Danish).

Johnson, P., 1998. Usability and mobility: interactions on the move. Proceedings of the First Workshop on Human–Computer Interaction with Mobile Devices.

Karat, C.M., Campbell, R., Fiegel, T., 1992. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. Proceedings of CHI'92, ACM, New York, pp. 397–404.

Kjeldskov, J., Graham, C., 2003. A Review of MobileHCI Research Methods. Proceedings of the 5th International Mobile HCI 2003 conference, Udine, Italy. Lecture Notes in Computer Science. Springer, Berlin.

Kjeldskov, J., Skov, M.B., 2003a. Evaluating the usability of a mobile collaborative system: exploring two different laboratory approaches. Proceedings of the 4th International Symposium on Collaborative Technologies and Systems 2003, Orlando, FL. SCS press, San Diego, pp. 134–141.

Kjeldskov, J., Skov M. B., 2003b. Creating a realistic laboratory setting: a comparative study of three think-aloud usability evaluations of a mobile system. Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction, Interact 2003. IOS Press, Zürich, pp. 663–670.

Koppinen, A., 1999. Design challenges of an in-car communication UI. Proceedings of the Second Workshop on Human Computer Interaction with Mobile Devices, Mobile HCI'1999, Edinburgh, Scotland.

Kristoffersen, S., Ljungberg, F., 1999. Mobile use of IT. In: Käkölä, T.K. (Ed.), Proceedings of the 22nd Information Systems Research Seminar in Scandinavia, Vol. 2, Department of Computer Science and Information Systems, University of Jyväskylä, pp. 271–284.

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, ACM, New York, pp. 206–212.

Luff, P., Heath, C., 1998. Mobility in collaboration. Proceedings of CSCW'98, ACM, New York, pp. 305–314.

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

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., Kirakowski, J., 1998. Comparative evaluation of usability tests. Proceedings of the Usability Professionals Association Conference, pp. 189–200.

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

Nielsen, C., 1998. Testing in the field. In: Werner, B. (Ed.), Proceedings of the third Asia Pacific Computer Human Interaction Conference, IEEE Computer Society.

Nielsen, J., 1993. Usability Engineering. Morgan Kaufmann, Los Altos, CA.

Nielsen, J., 2000. Why you only need to test with 5 users, Alertbox, http://www.useit.com/alertbox/20000319.html.

Pascoe, J., Ryan, N., Morse, D., 2000. Using while moving: HCI issues in fieldwork environments. Transactions on Computer–Human Interaction 7 (3), 417–437.

Petrie, H., Johnson V., Furner, S., Strothotte, T., 1998. Design lifecycles and wearable computers for users with disabilities. In: Proceedings of the First Workshop on Human–Computer Interaction with Mobile Devices, Glasgow.

Pirhonen, A., Brewster, S.A., Holguin, C., 2002. Gestural and audio metaphors as a means of control for mobile devices. Proceedings of CHI'2002. ACM, New York.

Preece, J., Roger, Y., Sharp, H., Benyon, D., Holland, S., Carey T., 1994. Human–Computer Interaction. Addison-Wesley, Reading, MA.

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.

Rubin, J., 1994. Handbook of Usability Testing. Wiley, New York.

Salvucci, D.A., 2001. Predicting the effects of in-car interfaces on driver behavior using a cognitive architecture. Proceedings of CHI'2001, New York, pp. 120–127.

Sawhney, N., Schmandt, C., 2000. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. Transactions on Computer-Human Interaction 7 (3), 353–383.

Sharples, M., Corlett, D., Westmancott, O., 2002. The design and implementation of a mobile learning resource. Personal and Ubiquitous Computing 6, 220–234.

Thomas, B., Grimmer, K., Zucco, J., Milanese, S., 2002. Where does the mouse go? An investigation into the placement of a body-attached TouchPad mouse for wearable computers. Personal and Ubiquitous Computing 6, 97–112.

Virzi, R.A., 1992. Refining the test phase of usability evaluation: how many subjects is enough? Human Factors 34 (4), 457–468.