Anne Kaikkonen, Aki Kekäläinen, Mikael Cankar, Titti Kallio, and Anu Kankainen. 2008. Will laboratory test results be valid in mobile contexts? In: Joanna Lumsden (editor). Handbook of Research on User Interface Design and Evaluation for Mobile Technology. Hershey, USA: Information Science Reference, volume II, chapter LIII, pages 897-909.

Handbook of Research on User Interface Design and Evaluation for Mobile Technology

Volume II

Joanna Lumsden
National Research Council of Canada
Institute for Information Technology – e-Business, Canada

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Copy Editor: Joy Langel, Katie Smalley, and Angela Thor
Typesetter: Jeff Ash
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

Library of Congress Cataloging-in-Publication Data

Handbook of research on user interface design and evaluation for mobile technology / Joanna Lumsden, editor.

p. cm.

Summary: "This book provides students, researchers, educators, and practitioners with a compendium of research on the key issues surrounding the design and evaluation of mobile user interfaces, such as the physical environment and social context in which a device is being used and the impact of multitasking behavior typically exhibited by mobile-device users"--Provided by publisher.

Includes bibliographical references and index.

ISBN 978-1-59904-871-0 (hardcover) -- ISBN 978-1-59904-872-7 (ebook)

1. Mobile computing--Handbooks, manuals, etc. 2. Human-computer interaction--Handbooks, manuals, etc. 3. User interfaces (Computer systems)--Handbooks, manuals, etc. I. Lumsden, Joanna.

QA76.59.H36 2008

004.165--dc22

2007024493

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter LIII
Will Laboratory Test Results be Valid in Mobile Contexts?

Anne Kaikkonen
Nokia, Finland

Aki Kekäläinen
TeliaSonera, Finland

Mikael Cankar
TeliaSonera, Finland

Titti Kallio
TeliaSonera, Finland

Anu Kankainen
Idean Research, Finland

ABSTRACT

The phenomena a usability test in the field reveals are different from those uncovered in a classical usability test conducted in a laboratory setting. Comparison studies show that these findings are more related to the user experience and user behaviour than usability and user interaction with the device. Testing in the field is a necessary part of the product development cycle, but the question is what and how to test. Duplicating a laboratory usability test method in the field may not make sense in many cases because the required extra effort does not result in comparable added value, as far as understanding user interaction. Studying user behaviour, on the other hand, requires a less controlled test setting.

INTRODUCTION

The mobile context challenges the user of a mobile system in many ways. The user's attention is divided between interaction with a mobile application and interaction with the environment and other people. The complexity of a real usage environment is a concern for usability practitioners. The question is: Can usability tests conducted in laboratory settings provide results that are valid in real-life mobile contexts?

In this chapter the benefits and drawbacks of mobile application usability testing in laboratory settings and in the field will be discussed. First, the latest views on the nature of the mobile context and how it challenges the mobile user will be presented. Then, some recent discussions concerning usability testing methods in general and issues regarding testing within industry vs. testing with academic goals will be described. After that, studies comparing usability testing in a laboratory with testing in the field, including this study, will be looked at. Some recommendations regarding when to test mobile applications in the laboratory and when in the field, also will be provided.

THE MOBILE CONTEXT

Usability practitioners talk of testing the usability of mobile applications in the field because laboratory settings differ from real usage environments. The mobile context is often considered to be too complex for laboratory simulation. To understand the background, that is, the different aspects of mobility, first the complexity of mobility as a concept needs to be discussed. In the following section, what kinds of challenges mobile users might face when using mobile devices and services on the move, will also be talked about.

Mobility is More than Just being on the Move

The simplest way to think about mobility would be to state that a mobile person is on the move. People travel from place A to place B, visit other places, and wander inside the places (Kristoffersen & Ljungberg, 1999). In reality, we need to remember that people also stop moving and "claim space" for their actions in mobile contexts. For example, people in a bus might pick up a newspaper for some privacy from the surrounding people, or a group of friends who happen to meet each other at a metro station gather in a circle to converse in private, as shown in an ethnographical study by Tamminen et al. (2004). In a sense, people can block out at least parts of their surroundings and concentrate on the task at hand.

Mobile device users may use their devices to build a private environment: When someone needs some privacy in the middle of a busy place, they can take their personal space with the device. A good example could be using a laptop to set up a temporary office, like a "nomadic tent," in a crowded cafe or an airport. The same thing can happen when using a mobile phone. It is not uncommon to see people in public transportation reading, sending messages, or engaging in other activities using their mobile phone. It is a way to gain some privacy.

Mobility on a Larger Scale

Being on the move and stopping to interact with a mobile device does not, however, convey the whole picture of mobility. Kakhira and Sørensen (2002) argue that mobility is not just being on the move but, far more importantly, related to the interaction between mobile people—the way in which people interact with each other in their social lives. Therefore, they suggest expanding the concept of mobility concept by three interrelated dimensions of human interaction: spatial, temporal, and contextual mobility.

Spatial mobility means that not only people, but also objects (such as a mobile phone), symbols (such as news through TV satellites), and spaces (such as virtual communities) move (Kakhira & Sørensen, 2002). Changes in physical contexts are not the only challenge for mobile users, but moving symbols and virtual spaces also require attention and special understanding. This is quite

often apparent in usability tests when the user needs to understand, for example, the billing model behind network services or the location of files in virtual spaces.

Temporal mobility is related to how mobile users perceive and use time. Mobile technologies may allow people to speed up time or save it according to their needs. The temporality of human interaction, however, can no longer be explained from a linear "clock-time" perspective alone; it now consists of multiple temporal modes based on each actor's perspective and interpretation of time itself. The increasing temporal mobilization of human interaction is creating new opportunities and constraints for the ecology of social life. (Kakhira & Sørensen, 2002).

Moreover, changing contexts (culture, language, non-verbal communication, environment, other devices, etc.) challenge the mobile user. Contexts in which people act continuously redefine both how they interact with others and with mobile devices. Mobile technologies allow people to interact with each other free from many contextual constrains (Kakhira & Sørensen, 2002). Spatial mobility, on the other hand, requires adapting to constantly changing contexts. For example, a location-based application can engage the mobile user much more in the current context than using e-mail, for example.

The Challenges of Mobility

As seen from the previous section, mobility is not a simple concept. Its complexity has an impact on how mobile users interact with mobile applications and how they experience these applications.

Active Participation

As Kakhira and Sørensen (2002) argue, mobility is frequently psychosocial in nature. Often, mobility also requires active participation (Oulasvirta et al., 2005), which might interfere with mobile applications interaction. An ethnomethodologically oriented mobility study in urban areas conducted by Tamminen et al. (2004) revealed four characteristics of mobile contexts, illustrating active participation.

The first characteristic was that people often have a plan when moving from one place to another, but the plan functions as a framework and leaves space for situational acts such as dropping by, ad hoc meetings, and other forms of "sidestepping." This requires flexibility from the plan and, to a degree, in navigation. It also means that some mobile interaction might be planned (such as reading e-mails on the train), but some of the interaction can consist of ad hoc activities.

The second characteristic the study revealed was that people on the move often solve their navigational problems by interacting with other people. People not only ask other people to give advice on routes or timetables, but they also inform other people about schedule changes, or negotiate what to do next. This is often done using mobile devices. The assumption is that navigational tasks may sometimes be of the highest priority when on the move and other mobile HCI tasks need to release resources for them.

The third characteristic seen in the study was that time plays a crucial role when moving in urban areas. It is often argued that mobile devices free people from the limitations of time and place. Nonetheless, when people are on the move in urban areas, they do face temporal tensions. Some situations may accelerate, so that hurrying and multitasking is necessary, tasks need to be prioritized and some tasks may have to be given up. Sometimes urban mobility requires slowing down or even stopping. For example, missing a bus means you have to wait for the next bus, or arriving early for a meeting, you have to wait for it to start. These temporal tensions may influence the cognitive resources available for interacting with mobile devices.

Finally, the fourth characteristic was that people have a need to multitask while on the move, but mobility may restrict it. For example, opening a door with a key while trying to talk on a mobile phone is challenging as is trying to listen to metro station announcements while talking on the phone. Sometimes, there might also be a need for multiple mobile HCI tasks such as writing a text message and using the calendar, which requires switching between different applications and orientations.

These multiple mobile HCI tasks are easy to simulate in laboratory settings, but tasks involving more than mobile devices are more difficult to simulate as part of a laboratory test environment.

The Competition for Cognitive Resources

Being aware of the environment and tasks related to navigation engages a big part of people's attention and cognitive resources when on the move. Tasks related to this context, like choosing the right bus or metro, or avoiding being hit by a car while crossing a street, are people's primary tasks in an urban environment. While people use all their senses to monitor what is going on around them, visual resources are particularly important for various tasks (as concluded by Lumsden & Brewster, 2003, among others). Estimating the arrival time of the metro, finding a seat, noticing a friend in the same compartment, and getting off the metro on the right station are just a few of the tasks that require visual cues. Using a mobile device or application competes for cognitive resources with the user's natural active participation with the environment and navigation tasks.

Gonzales et al. (2004) studied how information workers manage multiple activities in a normal work environment and were surprised by the high level of discontinuity in the execution of daily tasks. People spend an average of 3 minutes working on any single task before switching to another. Given their limited cognitive resources, it is interesting that people manage these streams at all.

People's attention is even more fragmented when they move around. According to Oulasvirta et al. (2005), this multitasking in the field leads to a depletion of resources available for task interaction and eventually results in the breakdown of fluent interaction. According to their field study, the test users' continuous attention on the mobile device fragmented and broke down to bursts of just 4 to 8 seconds. The users' attention was diverted from the mobile device to the surrounding environment up to eight times during the time it took for a single mobile Web page to load.

USABILITY TESTING

Since the mobile context challenges the user in so many ways, it is understandable that the ecological validity of usability studies has been a hot topic since the early days of mobile devices. This section will review studies that attempt to resolve the differences between testing mobile applications in field and laboratory settings. In order to discuss the differences, first, what is being talked about needs to be defined.

The Principles of Usability Testing

Usability testing based on the thinking-aloud protocol was originally created and presented by K. A. Ericsson and H. A. Simon in their article Verbal Reports and Data (1980) and a follow-up book, Protocol Analysis (1984). The goal of a test is to study end-user *behaviour* regarding the use of an application or a service, not the user's *opinions*. Questions about opinions can be included as a part of a test session, but basically the usability test method is concerned with observing user behaviour and how the user interface of an application matches the human way of thinking and acting. The usability test protocol is described in more detail in several handbooks, for example in Rubin's (1994) Handbook of Usability Testing.

In a usability test, a test user is advised to think aloud while s/he tries to accomplish a given task. The user is asked, for example, to find a piece of information using a search engine. By observing and listening to the user, the facilitator and other observers find out how the user's thinking proceeds and what s/he expects to find in the user interface. All the silent moments in a session, the wrong paths the user chooses, questions, and so forth, indicate problems in the user interface structure, terminology, or navigation.

Usability testing is a qualitative method as opposed to a quantitative method (questionnaires with statistical analysis, etc.). When the need to collect users' opinions arises, other methods such as those used for market research, must be used. In these studies the number of respondents is typically far higher than in usability tests, that is, up to hundreds or even thousands of people.

Usability testing conducted during an industrial product development process is usually not academic research: The goal of a usability test is to improve the system being developed. Sufficient results are often achieved with 5 to 10 users per test iteration, although all problems may not be detected. The goal of academic HCI research is to better understand users' behaviour and interaction models, as well as improve the methods used in product development. In order for the results to be reliable and to help comparison between studies, the number of test users should be higher. In a paper by Faulkner (2003), a minimum of 95% of usability problems were found with 20 users and variation between groups was fairly small.

Usability Testing in Industry

Usability testing can be adapted for different applications. Resources for application development in an industrial context are usually limited, and usability activities such as user-centred design and usability testing, must be performed cost-effectively. The goal of usability testing in product development is to find severe and disturbing usability problems within the strict limitations of project budgets and deadlines. It is rarely possible, or necessary, to remove every minor glitch from a user interface before a product launch.

Since time and resources are critical, companies look into the most efficient ways to find usability problems in products. Sometimes this means taking shortcuts that should not be taken. Indeed, Ramey and Boren (2000) have investigated the practice of testing and found that often the original usability testing procedure is not properly followed. When resources are limited, attention must be paid to expertise in testing. There usually is not much time for trial and error or training.

Wixon (2003) has also raised special issues to take into consideration when testing cost-effectively in the business world. He says that it is not just the usability problems found in the tests that are relevant for product development, but it is necessary to use a testing framework, which defines how the service can be improved in the shortest time with the least effort.

A commonly accepted recommendation in industry environments is that usability tests with only five test users can reveal 85% of user interface problems (basic human cognitive processes vary little) (Nielsen, 2000). Such a requirement can be criticized or evaluated further in an academic sense, but it is a good example of the efficiency demands present in product development projects.

In addition, usability testing in the field is more time consuming than laboratory testing (Kaikkonen et al., 2005; Kjeldskov & Graham, 2003). Without concrete proof to support the theory that testing in a real-life context is significantly better than a laboratory test, companies have good reason to question whether investing in more expensive and more time consuming field tests is worthwhile.

Usability Testing in the Field and in the Laboratory

Since the advent of mobile systems and services, usability practitioners have discussed the ecological validity of laboratory usability studies and how much results could be improved by testing in the field. A controlled environment is far removed from real-life contexts and may lead to biases in test results. Maintaining dedicated usability laboratories is an expense for companies and their very need has also been questioned.

Modern Testing Equipment

It is only recently that truly mobile recording equipment environments have become available to researchers. Unsurprisingly, up to recent times most (71%) mobile device evaluations have taken place in laboratory settings (Kjeldskov & Graham, 2003). Without proper equipment it was impossible to gather field test data the way it is done in a laboratory environment, that is, by following the user's actions step by step and recording them for further analysis. Miniature cameras now allow proper data gathering in a variety of test settings without obstructing the user in his/her performance of the tasks. The dynamics between the moderator and the user can be similar to those found in a laboratory environment.

Once miniature cameras became available, researchers have used them in different ways to study users and services in real-life environments. Roto et al. (2004) show how usability experiments can be conducted in a field environment using a mobile miniature camera for recording not only the user's actions on the mobile device, but also the user's surroundings. They recommend that field usability tests be conducted in situations where user interaction with the environment is investigated, in addition to interaction with the system.

Comparative Field and Laboratory Studies

As technology allowed usability testing to move out of the laboratory, researchers started studying what this meant for their studies. The following will present a cross-section of recent research into the differences of testing in the laboratory and in real-life environments. The papers tend not to define what is meant by the terms *usability* and *usability problem*, making it very difficult to understand how the outcomes actually differ from each other and whether the authors are talking about the same issues in their conclusions. Also, the number of test users is often so small that the variation of findings within a group is likely to be as big as the variation between the groups.

Kjeldskov et al. (2004) conducted a comparison study of an expert application for health care professionals. The study was conducted in a laboratory setting that was built to resemble a part of a physical space in a hospital department. The tasks in the study were related to the daily activities of the hospital personnel. The field test was conducted in an actual hospital environment. In the study, with six test users in the field and six test users in the laboratory, Kjeldskov et al. came to the conclusion that testing in the field adds little value compared to the laboratory test. Molich's classification was used in the analysis, but no clear definition of a usability problem was given. Some problems in this study did not come out in the field setting and the field setting involved events that decreased control over the study.

Duh et al. (2006) also conducted a comparison study between field and laboratory settings. Twenty users participated in this test, 10 in both settings. The laboratory part was conducted at the usability laboratory of a university and the field test was conducted on a train in Singapore. The tasks in the study were related to activities people might engage in while using public transportation, such as normal use of a mobile phone. In the laboratory, the researchers gave the test participants task scenarios to help them understand the actual use context. In this study more critical problems were found in the field test setting than in the laboratory. The reasons for this, according to the researchers, were that different disturbances (noise, movements, lack of privacy), among other factors, affected test user performance. They concluded that some problems are only found in a field environment. Comparison of these results to other studies is difficult as no explicit definition of what constitutes a usability problem is provided. The used definition seems to differ from standards or other studies in a sense that it seems to include users' behavioural patterns and user interaction with the environment.

Holtz Betiol and de Abreu Cybis (2005) performed a comparison test of three approaches: laboratory tests with a PC-based emulator, laboratory tests with an actual mobile device, and field tests. They had groups of 12 users in each case and the tasks the users performed were related to the use of a mobile portal, that is, tasks that are relatively common for ordinary users. The field part of the test was not performed on the move, but participants were placed in a noisy environment. The study is noteworthy because it actually defined what is meant by a usability problem: the authors used the ISO 9241-11 definition. The results in this study did not show statistically significant differences between the laboratory and field tests when the mobile device itself was used.

Baillie and Schatz (2005) explored multimodality in two conditions: they used testing as one part of the application development process. They used a fairly small number of four to six users per setting in their experiment since the test was only one of several methods used to evaluate the

system. The researchers were surprised by the results. It took less time to complete the tasks in the field than in the laboratory. More problems were found in the laboratory environment than in the field, even though there was no difference between the environments when it came to critical problems. Again, the definitions for usability and usability problems were not provided in the paper, making it difficult to compare the results with other studies.

Our Study

The authors of this chapter work closely with mobile terminals and services. The details of testing methods are critical to the work. Based on a number of usability tests that were ran in the work, both in the laboratory and in various out-of-laboratory environments, it was felt that the experiences differed somewhat from ones detailed in the studies that were being read. To understand and verify whether testing in the two environments produces different results, two parallel usability evaluations of a mobile application were organised. One test took place in a typical usability laboratory and the other in the field, including tasks like walking in a shopping centre and using the subway (Kaikkonen et al., 2005).

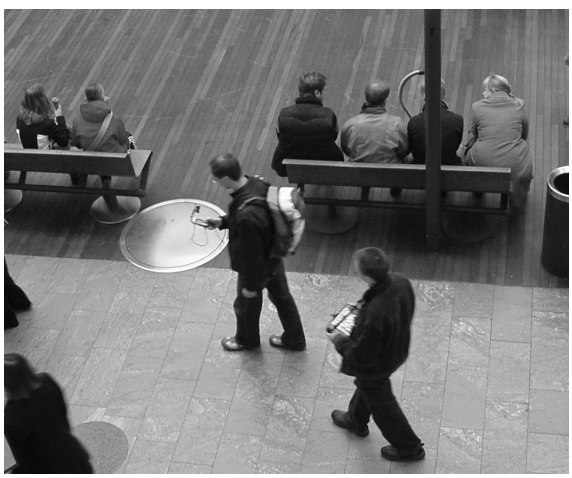
The two test rounds in different contexts were designed to be as similar as possible. The thinking-aloud protocol and the same predefined series of test tasks were used in both cases—the goal was to make sure the context was the only changing variable. The users were prompted to explain what they were doing, what they expected to happen when making selections, and whether something unexpected happened after the selection. One can question whether conducting the same test in such different environments is meaningful, but it was wanted that the test situations be made as similar as possible to find out if changing one variable, the environment, would make a difference.

The total number of test participants was 40, 20 in both settings. Using a relatively large number of participants meant that variations within groups should not be bigger than between groups.

Special equipment allowed the moderators to run and record the tests. A test user carried a backpack with miniature cameras for recording both the mobile device interface and the user (facial expressions). The moderator could follow the camera image, live, from a wireless six inch monitor that also had an additional camera to record the user's surroundings or anything the moderator considered relevant. The test situation can be seen in Figure 1.

The problems found in different test settings were listed and analyzed both quantitatively and qualitatively. This study produced several results, some of which were contrary to the expectations.

Figure 1. Field test ongoing



The main finding was that there was no difference in the number of problems occurring in the two test settings. In fact, the same usability problems were uncovered in both test settings.

Some differences could, however, be seen between the two settings when the frequency of each problem was studied. The problems that occurred more often in the field seemed to be related to understanding the logic of the relatively complicated application. On the other hand, there were also complex issues where no difference between the two test settings could be discerned.

Even individual task execution times in the field were no longer than those in the laboratory. However, the total time needed for the testing was longer in the field because of the preparation necessary for using the field-testing equipment. Some of the test tasks were performed at specific locations to make the tasks more sensible (for example, taking a picture of flowers in a shopping centre), which also contributed to the longer time the test took in the field.

Interesting observations were made about user behaviour that was more related to user experience than usability during the tests. The many interruptions in the field test did not seem to affect the user's performance. Other metro passengers, for example, did not seem to bother users, even if they came to talk to the moderator. In an extreme case, four security guards at a shopping centre were staring at a user whose backpack looked very suspicious, but the user did not even notice the guards. This was not the case with the moderator, who felt quite awkward at the time.

The users in the field tests concentrated intensely on the tasks at hand, and a few users were able to perform all the tasks while walking. Given a more complex task, the users sought a spot where they were safe from surrounding disturbances, essentially creating a bubble of privacy around them. Creating a safe haven in a public place is natural for users, but how much of an impact the artificial nature of the test setup had on the users' cognitive load also has to be considered. The users typically did not have access to their own address books and had only limited experience with using the device being tested. These issues may mean that users are not able to multitask as well as they would when using their own, familiar devices.

Slowing down or even stopping to perform complex tasks is very much in line with the findings of Mizobuchi et al. (2005), who, in a controlled environment, observed that the walking speed of test participants was fairly slow when typing on the mobile phone. This behaviour gave insight into the difficulty level of the tasks and is difficult to observe in a laboratory setting.

SUGGESTIONS FOR FIELD TESTING

Most of the comparison studies presented in the previous section, including the one we conducted, indicate that conducting usability studies of mobile user interfaces in the field is not worth it. In some cases, however, it may make sense to conduct field tests depending on what kinds of user interfaces are being tested and what kinds of usability problems are to be expected. For example, if the intention is to test talking on the phone in noisy environments such as the metro (Duh et al., 2006) and it's not possible to realistically simulate the noise in a laboratory, it makes sense to test in the field.

Location-based and context-aware services are another example. Testing whether people can find the right route using a GPS navigation tool in a laboratory would be difficult, as this depends on how the user succeeds in transferring the map representation to the actual environment.

Tactile feedback is another area that is difficult to study in the laboratory. The difference in the user's attention level may also have implications for how they notice progress indicators in the application. In a real environment users may pick up a newspaper or check their own phone for calls while waiting for the device to finish a download task, whereas in the laboratory they just stare at the phone screen for minutes at a time (Kaikkonen et al., 2005).

Sometimes usability testing requires little additional effort as part of a field trial that is already taking place. During such trials, prototypes of a system being designed are given to test users for

use in their everyday life for a longer period of time (such as 4 weeks). During this time the users can be interviewed and observed several times in order to study not only usability issues but also behavioural patterns emerging from interaction with the prototype. Log files can be used to collect additional user data. The result is a deeper understanding of why, how, and in what contexts users would use the system being developed (see for example, Mäkelä et al., 2000). When prototypes are tested with groups of people who interact with each other during the trial, social interactions can also be studied. As Kakhira and Sørensen (2002) point out, mobility also involves social interaction and not only being on the move.

Choosing a Location for Out-of-Laboratory Testing

Mobile phones and other mobile devices are used 'anywhere' and defining a good out-of-laboratory location to test a device or an application is not a simple task. The location used in the test should be one where people normally use mobile devices. Specifically, it should be socially acceptable to use such a device at the test location.

Calling on the mobile phone, for example, may irritate bystanders (Love & Perry, 2004), but there are places where even text messaging is inappropriate. Test users are usually acutely aware of social norms related to phone usage in public places in their own environment and breaking the norms might make the test users feel uncomfortable, like Palen et al. reported in their study (2000). Users should not be given tasks that force them to act against the social norms of the test location.

Diverse places such as cafés, cinemas, transportation, and streets, have different social codes, depending on how they are built. Fyfe (1998) writes about the effect of architecture on people's behaviour in public places and differences in different cities; the way the environment is built either encourages or discourages social communication, walking in the streets, and other behaviour. These kinds of architectural effects need to be taken into consideration when planning the test environment.

When testing in an unfamiliar environment, it would be beneficial to ask local people about norms and social codes, or observe how people behave prior to test planning. This also helps evaluate the validity of the test results. In order to understand how ecologically valid the test situation is, user behaviour related to the test device and service needs to be analysed. And even that is not enough. When running a test in a public place, whether user behaviour differs from the social code in that particular environment also needs to be observed. The ways in which people generally create private spaces in public areas should be understood in order to draw the right conclusions from a test user's behaviour during a test session (Kopomaa 2000).

If a test is conducted in the "wrong" place, the results may give more insight on test user interaction with the environment and other people than with the tested device or service. Testing in a socially unacceptable place may also create unnecessary stress for the user and s/he may not be able to concentrate on other issues.

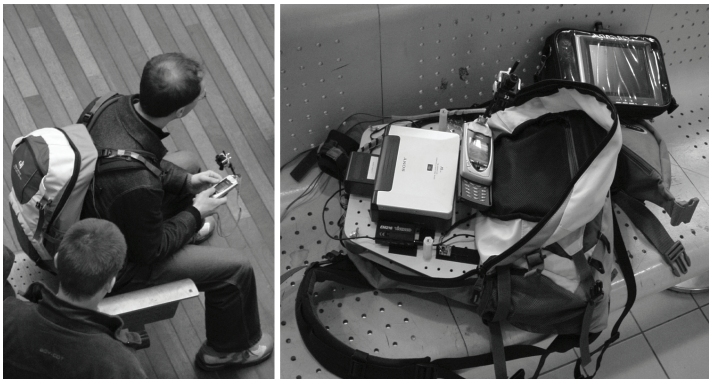
Choosing a test location may also depend on what usage is studied: the initial experience of learning to use a device or later, continued usage. Based on the information with mobile phones and services, for example, people tend to try out new gadgets at home or some other peaceful place, while the eventual usage environment may well be a bus or a crowded restaurant.

The Logistics of Field Testing

The relatively complex equipment necessary for recording user interaction in the field requires more preparations than the familiar equipment used in a laboratory. There are batteries to recharge, the backpack must be adjusted for each user, and explaining how the user should behave during the test typically takes longer. The complexity of the equipment can be seen in Figure 2.

This means that field tests take more time than laboratory tests, as can be seen in table 1 (according to our study). In practice, one can run fewer tests per day in the field than in the laboratory.

Figure 2. Equipment used in field tests



Field tests are vulnerable to unexpected events, such as rain or bus schedules. These risks should be listed before the test is run with actual test users. Since the environment cannot be controlled in the same way as the laboratory, the researchers should also have a backup plan or recruit an extra user, just in case. Running a pre-test or a pilot is critical to the success of a field study. This helps to reduce the risks due to the technology used, but it also helps identify factors that may influence the analysis of the results. If the user moves around during the test, for example, is there a location where the lighting makes it impossible to see the text on a screen, or the surrounding noise blocks out the notifications of the device? If the test focuses on software rather than hardware issues, these kinds of environmental disturbances may make it impossible to get any meaningful results from the test.

There are several test planning issues that must be specified in greater detail for a field test than a laboratory test—particularly if multiple moderators run the test or the tests are outsourced. Examples of these issues are moderator prompting, timing between questions, how to react to external interruptions, and to what extent test user behaviour is controlled. Since the field setting is less predictable, specifying these details takes additional effort.

It is important to be open about the nature of the test when recruiting users. Some users may not be willing to participate when they hear the test will take place in a public location—it happened with a few users. Facing this issue while

Table 1. Differences between locations

	Laboratory	Field
Total test time per user, average	35 min.	45 min.
Instructions and preparations per user, estimated time	10 min.	20 min.
All user interface problems found	Yes	Yes
Users easily understood the application concept	Yes	No
User behaviour can be observed in a natural environment	No	Yes
Environment can be fully controlled	Yes	No
Suitable for usability testing	Yes	Yes
Suitable for testing a concept or service idea	With restrictions	Yes

recruiting is a lot easier than having irate users quitting in the middle of the test.

The effect of the recording equipment on the test user needs to be taken into consideration. Even with miniature cameras, the backpack may be too heavy for some users, possibly limiting the duration of the test sessions. Having the moderator carry as much equipment as possible on behalf of the test user is recommended. If the equipment is conspicuous, the test user may find carrying it embarrassing which may produce a bias in the test results and make it harder to recruit users.

The tasks planned for the field test need to be natural for the test environment. As discussed earlier, the environment and an unfamiliar device increase the users' cognitive load and they may not be able to multitask as well as they would in a normal situation, using a familiar device.

Even after careful preparation, field tests are unique events. Potential interruptions and overall user behaviour need to be taken into consideration when analysing the data from each test.

CONCLUSION

Most comparison studies, including this study, indicate that conducting usability studies of mobile applications or devices in the field in order to find usability problems alone in user interaction with a system, that is, usability problems as defined in the ISO 9241-11 standard, is not worth it. Based on these findings, the recommendation for most testing needs is to use the available resources to perform several quick laboratory tests iteratively during the design process, rather than concentrate efforts on a single field test.

There are also situations where laboratory testing is not enough. In some cases, the limitations of a laboratory setting may be technical. GPS navigation, for example, does not work indoors. Some environmental factors, such as noise, can be difficult to simulate realistically. In other cases the limitation may be a result of how the device is used together with the environment. Again using GPS navigation as an example, a real test task involves mapping information from the device to the surroundings.

Field-testing can also be useful when the purpose is, in addition to testing the usability of a user interface, to gain knowledge about user behaviour in a natural environment, that is, to understand where users might use the service. During the first stages of the product development process, the most important information comes from understanding users and the environments where the service is going to be used. Observation, in-depth interviews, and other methods used in psychology and sociology provide information that better describes the needs of the users, as well as possibilities and restrictions for the service. From a service design point of view, if a service is supposed to be usable while the user is on the move, the designer has to know what "on the move" means for the users of that particular application. In general, it is crucial for product developers to understand the users' usage patterns and multitasking requirements because it helps create better services.

Later on, with prototypes or first versions of the service, evaluation comes in to play, but there are a lot of unanswered questions beyond simple usability problems. Conducting a usability test in the field is one way to find usability problems and get information that can be acquired more easily in the right context. On the other hand, user testing as part of a product development process does differ from user research, even if the methods used can be similar.

Finally, one explanation for similar findings in laboratory and field environments could be that most mobile services require such a high level of concentration, forcing users to create "a bubble" around them and stop other activities. Maybe user interfaces that are easier to use will not only open up new possibilities for using the services in different situations, leading to a better user experience, but also bring opportunities for mobile services and device manufacturers. This would also mean that testing in the field should be re-evaluated if new, easier user interaction models change user behaviour on the move.

REFERENCES

Baillie, L., & Schatz, R. (2005). Exploring multimodality in the laboratory and the field. In proceedings *In Proceedings of ICMI 2005.*

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261-278.

Duh, H. B.-L., Tan G., & Chen, V. (2006). Usability evaluation for mobile device: A comparison of laboratory and field tests. In *Proceedings of MobileHCI 2006*

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215-251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments and Computers*, *35*(3), 379- 383.

Fyfe, N. R. (1998). *Images of the street* (Introduction, pp. 1-10). NY: Routledge.

González, V. M., & Mark, G. (2004). Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings CHI'04* (pp. 113-120). ACM Press.

Holtz Betiol, A., & de Abreu Cybis, W. (2005). Usability testing of mobile devices: A comparison of three approaches. In *Proceedings of Interact 2005.*

ISO 9241-11 (1998). *Ergonomic requirements for office work with visual display terminals (VDT)s – Part 11: Guidance on usability.*

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio T., & Kankainen A. (2005). Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, *1*(1), 4-16.

Kakhira, M., & Sørensen, C. (2002, January 7-10). Mobility: An extended perspective. In *Proceedings of the Hawaii International Conference on System Sciences*, Big Island, Hawaii.

Kjeldskov, J., & Graham, C. A. (2003). Review of mobile HCI research methods. In *Proceedings of the 5th International Mobile HCI 2003 Conference*, Udine, Italy. Sringer-Verlag, LNCS.

Kjeldskov J., Skov, M. B., Als, B. S., & Høegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Proceedings of the 6th International Mobile HCI 2004 Conference* (pp. 61-73). Glasgow, Scotland. Berlin: Springer-Verlag.

Kopomaa, T. (2000). *City in your pocket. Birth of the mobile information society*. Helsinki: Gaudeamus.

Kristoffersen, S., & Ljungberg, F. (1999). Mobile use of IT. In T. K. Käkölä (Ed.), *Proceedings of the 22nd Information Systems Research Seminar in Scandinavia Conference* (Vol. 2, pp. 271-284).

Love S., & Perry, M. (2004). Dealing with mobile conversations in public places: Some implications for the design of socially intrusive technologies. In *Proceedings of CHI 2004*. Vienna: ACM.

Lumsden, J., & Brewster, S. (2003). Paradigm shift: Alternative interaction techniques for use with mobile & wearable devices. In *Proceedings of CASCON'03.*

Mizobuchi S., Chignell, M., & Newton, D. (2005). Mobile text entry: Relationship between walking speed and text input task difficulty. In *Proceedings of the MobileHCI 2005.* Salzburg, Austria: ACM.

Mäkelä, A., Giller, V., Tscheligi, M., & Sefelin, R. (2000). Joking, storytelling, artsharing, expressing affection: A field trial of how children and their social network communicate with digital images in leisure time. *In Proceedings of CHI 2000* (pp. 548-555).

Nielsen, J. (2000). *Why you only need to test with 5 users*. Retrieved January 15, 2007, from http://www.useit.com/alertbox/20000319.html

Oulasvirta, A., Tamminen, S., Roto, V., & Kourelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. In *Proceedings of SIGCHI Conference on Human factors in Computing Sytemes* (pp. 919-928). NewYork: ACM Press.

Palen L., Salzman M., & Youngs, E. (2000). Going wireless: Behavior & practice of new mobile phone users. In *Proceedings of CSCW*. Philadelphia: ACM.

Roto, V., Oulasvirta, A., Haikarainen, T., Lehmuskallio, H., & Nyyssönen, T. (2004, August 13). *Examining mobile phone use in the wild with quasi-experimentation* (HIIT Tech. Rep. 2004-1).

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

Tamminen, S., Oulasvirta, A., Toiskallio, K., & Kankainen, A. (2004). Understanding mobile contexts. *Personal and Ubiqitous Computing*, *8*, 135-143.

Wixon, D. (2003). Evaluating usability methods. *Interactions*, *July-August*, 29-34.

KEY TERMS

Field Test: Usability test in a real-life context

Laboratory Test: Usability test in a controlled environment

Usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use (ISO 9241-11).

Usability Laboratory: A controlled environment where evaluators set up usability tests and other experiments. In a usability laboratory all factors of the tested system can be controlled and high-quality data collection (video, etc.) is possible.

Usability Problem: Problems that influence the effective, efficient, and satisfactory use of the system in a specified context of use (ISO 9241-11)

Usability Test: User test that tests the effectiveness, efficiency, and satisfaction of the system in a specified context of use (ISO 9241-11) using known usability test protocols

User Experience: User's holistic experience with the product-user experience is an intra-user event which is the consequence of how well the product matches with user expectations, how well it supports the activities in different physical and social contexts. The entire user experience may not be possible to detect by using usability testing alone.