HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

EXAMINING MOBILE PHONE USE IN THE WILD WITH QUASI-EXPERIMENTATION

Virpi Roto, Antti Oulasvirta, Tuulia Haikarainen,
Jaana Kuorelahti, Harri Lehmuskallio, Tuomo Nyyssönen

August 13, 2004

HIIT
TECHNICAL
REPORT
2004-1

Examining Mobile Phone Use in the Wild with Quasi-Experimentation

Virpi Roto[1], Antti Oulasvirta[2], Tuulia Haikarainen[3], Jaana Kuorelahti[3], Harri Lehmuskallio[3], Tuomo Nyyssönen[1]

[1]Nokia Research Center
P.O. Box 407, 00045 Nokia Group, Finland
[virpi.roto, tuomo.nyyssonen]@nokia.com

[2]Helsinki Institute for Information Technology
P.O. Box 9800, 02015 HUT, Finland
oulasvir@hiit.fi

[3]Helsinki University of Technology
P.O. Box 1000, 02015 HUT, Finland

HIIT Technical Reports 2004-1
ISSN 1458-9451

Copyright © 2004 held by the authors.

Notice: The HIIT Technical Reports series is intended for rapid dissemination of articles and papers by HIIT authors. Some of them will be published also elsewhere.

EXAMINING MOBILE PHONE USE IN THE WILD WITH QUASI-EXPERIMENTATION

Virpi Roto[1], Antti Oulasvirta[2], Tuulia Haikarainen[3], Jaana Kuorelahti[3], Harri Lehmuskallio[3], Tuomo Nyyssönen[1]

[1]Nokia Research Center, Helsinki, Finland
[2]Helsinki Institute of Information Technology, Helsinki, Finland
[3]Helsinki University of Technology, Helsinki, Finland

Abstract

While increasingly better tools are available for examining human-computer interaction in a laboratory environment, we are only just starting to develop the methods and appropriate portable equipment within the mobile context. One problematic issue has been that current methods are not suited for attempting to establish causal relationships between context variables and interaction. In this paper we describe an experimental method, called quasi-experiment, and apparatus for recording mobile phone usage and the environment in a mobile context. Quasi-experimentation is based on 1) the best possible control over nuisance variables in the mobile test environment and 2) recordings of the user, interaction with the device, and the environment. It requires changes in the way we design, prepare, implement, and analyze interaction experiments. We learned that conducting a quasi-experiment is laborious without special tools that would decrease the amount of manual work. Quasi-experimentation is a promising investigation and evaluation methodology for the developers of mobile computing systems and applications.

1 Introduction

Human-computer interaction (HCI) and user behavior has been studied for decades in the laboratory environment and it is very clear by now how to conduct scientifically reliable experiments in controlled conditions. As the mobility of people increases and people start to use mobile and wearable devices for purposes similar to those of the office computer, we need to start examining how the mobile context affects the old rules of human behavior and the guidelines for building applications. The question is, how to examine user behavior and HCI in the (urban) mobile context when there are lots of uncontrolled variables in the environment, and with the subject moving among other people in public places.

In this paper, we investigate the possibility of taking laboratory experiments "into the wild". Experiments in the laboratory have been considered impossible to carry out in real-life contexts, because of poor control over possible nuisance factors. Our contribution lies in lessons learned about how to design, prepare, and implement field experiments, and how to prepare, code and analyze their data. We call this type of mobile field experimenting *quasi-experiments*. In aiming for stricter control, and therefore more valid inferences over causal relationships between the studied variables, several novel measures have to be taken, some of which this paper aims to reveal.

In this study, our task was to examine user's attentional resources while traveling around the city of Helsinki and using a mobile phone for Web-browsing at the same time. We wanted to find out how the users' attention was divided between the browser and the environment in different types of mobile contexts when the response times were long. We needed to follow the phone usage as well as user's eyes and the environmental conditions. It was clear from the beginning that there are no tools available to log all needed events automatically, so we had to manually record user behavior and the specified environmental conditions. Because the moderator could not constantly observe the phone display or the eyes of the user nor make detailed notes while on the move, we decided to use miniature cameras to videotape a) the phone display and keys, b) the user's face, and c) the environment. The videos were combined into one and analyzed by hand after the experiment.

This paper is about presenting and evaluating the method. The results that relate to attention resources in the mobile context will be presented at another time. We will first look at the related research that examines user behavior in the mobile context and then describe and analyze the method and apparatus used in this experiment.

2 Related Research

2.1 Methods

In order to get relevant answers to research questions, it is important to pay attention to the variables by which the data is gathered and analyzed. An experiment made by Hoyoung et al. focused on the use contexts of the Mobile Internet and their impact on usage patterns and usability problems [6]. They divided contextual information into eight categories: goal, emotion, hand, leg, visual distraction, auditory distraction, co-location, and interaction. Hoyoung et al. found that the movement of legs (move/stop) has a significant impact on the

usage of the Mobile Internet and on the kinds of usability problems. The eight categories provided us a good start, although the list did not examine e.g. temporal tension [11] that we think has an important effect on user's behavior in mobile context.

Several different methods have been used to gather data about human-computer interaction during user trials. Below is an analysis about their suitability for a study like ours.

Interviews can be conducted after a trial to gather participants' ideas on the interaction with the device. However, as has been known in experimental psychology for decades, interviews rely on memory and are not suitable to study micro-level interactions. The *Experience Sampling Method* (ESM) [2] tries to tackle this issue. It is based on the idea of randomly or semi-randomly "sampling" user experience, usually by "beeping" the participant and asking her to respond to a questionnaire (e.g., reporting the current task and if she feels she has been interrupted) on a sampling device, perhaps a PDA. This method, however, is annoying for the participant, suffers from missing data, and relies heavily on self-reports.

Diary studies of mobile interaction tend to concentrate on experiences, journeys, and important events during interaction, rather than interaction itself, but can also be used to study interaction in more detail, such as in a study by Czerwinski and colleagues [4]. They asked participants to log every task switch they made in an office to an Excel sheet and to rate the experienced difficulty of the switch. In the mobile context, this could be done similarly with a PDA, for example. However, the method does not avoid the shortcomings of ESM. Quite the contrary, as an additional demand is posed to the participant who must remember to fill the diary and interrupt her primary work for that. None of the methods can be used to study interaction at the micro-level, because keeping a log/diary at the same time while using the device is simply impossible.

Automatic logging escapes this criticism by recording interaction events without the user's notice. It is often done in order to capture user interaction during a field trial (e.g. [1, 12]). This, by itself, is severely limited in how well it can capture the use context, however. Although by itself it is often insufficient, automatic logging offers a reliable method for gathering interaction data without relying on subjective reports of the participant.

Observational studies, where the researcher follows the user with a video camera, can potentially also record aspects of the context [8]. Here, gathering data unobtrusively but comprehensively becomes the main problem. In an office environment cameras and other sensors can be hidden discreetly [5]. However, in the mobile context, moving the cameras along with the person while still being able to capture the important aspects of interaction and the context, is difficult. In *pair observations* [7] or "experience clips", a friend of the participant takes pictures or video clips during interaction, which makes the situation more relaxed. However, as is known with mobile images, people tend to fake and set up situations for imaging in a post hoc manner. In addition, the data is selected by the individual and her situationally arising interests—and are not determined by the research goals.

None of the methods studied thus far attempts anything like the experimental control pursued in laboratory experiments in HCI that has been inspired by experimental psychology. As we will argue in this paper, control in mobile contexts is difficult, but can be dealt with, to some extent. We will show that quasi-experimentation can overcome some of the limitations of other methods without sacrificing control too much.

2.2 Equipment

The equipment used to observe the user and the environment plays an important role in this type of study, so let us take a look at the apparatus used in similar mobile tests.

Woodruff et al. examined visitor behavior in a historic house while using an electronic guidebook prototype [12]. When using the electronic guide, "the visitors comments and conversations were recorded using wireless microphones, the visitors were videotaped by a camera placed in a corner of each room, the visitors were directly observed by the research escort, and the visitors' actions in the electronic guidebook were logged by the device for future reference." [12 p. 439] The apparatus in this study is interesting, especially the usage of the wireless microphone. If the study was on the street, the cameras could hardly be located in fixed positions but should be carried along. Thus, the equipment setup cannot be taken into use in truly mobile studies.

Cheverst et al. conducted an extensive field trial of a location-based tourist guide prototype in the city of Lancaster [1]. The goal of this study was to find the usability problems in the application, not to examine the behavior of the user, so they did not record the user's face or the environment. Instead, they used direct observation and logging for examining interaction between the user and the system.

In an experiment made by Oviatt et al. [10], the user actions were observed using a miniature video camera to record sound and events on the portable computer's screen. The devices were attached to the user's front and back waist bags to allow freedom of movement. An observer could view the same data on his own observation station and was able to intervene whenever needed. The observation station had to be relatively near to the subject, so the users executed the given tasks either in a quiet room or in a public cafeteria. As with Cheverst's study, Oviatt et al. did not record the environment nor the user's face.

Lyons & Startner's [9] recording system was improved in many ways. In addition to the mobile device display, they did record the environment with extra cameras and a microphone. To allow the movement and the recording of the actions in truly mobile context, the system was built into an ordinary vest. Instead of combining the different video streams into one, there were two camcorders in the pockets of the vest. Viewing the two videotapes in sync required a special viewing application. When more video camera pictures are needed, a quad setup will be necessary.

3 Method

Quasi-experiment is a type of quantitative research design conducted to explain relationships and clarify their nature [3]. Its main purpose is to examine causality in situations where complete control is not possible. Consequently, it has been used mainly in education, nursing, and medical research, where control of nuisance variables is not possible for ethical or practical reasons. Similarly, in conducting experiments "in the wild", i.e. mobile contexts, experimenters do not have total control over the events that take place in the experiment locations. However, quasi-experiments are *experiments*, rather than observations, because they test hypotheses and try to control as many threats to validity as possible in the wild.

While moving around the city, the subject executed Web-browsing tasks as autonomously as possible. The test moderator shadowed the user, gave the task instructions, and helped in

technical or route problems. The moderator was instructed to speak with the user only in the cafeteria.

We wanted to compare user behavior in the mobile context to the laboratory context, so either before or after the city tour, the subject executed four tasks in a laboratory.

Due to the large number of subjects, 32, we divided the sessions to 5 moderators. This meant we had to instruct the sessions carefully on paper.

3.1 Recording Setup

Our goal was to record the user's actions, the environment, and the state of the mobile phone for the whole 1.5-hour city tour. This was a challenging task and we needed a complex apparatus to fulfill the recording needs (Figure 1).

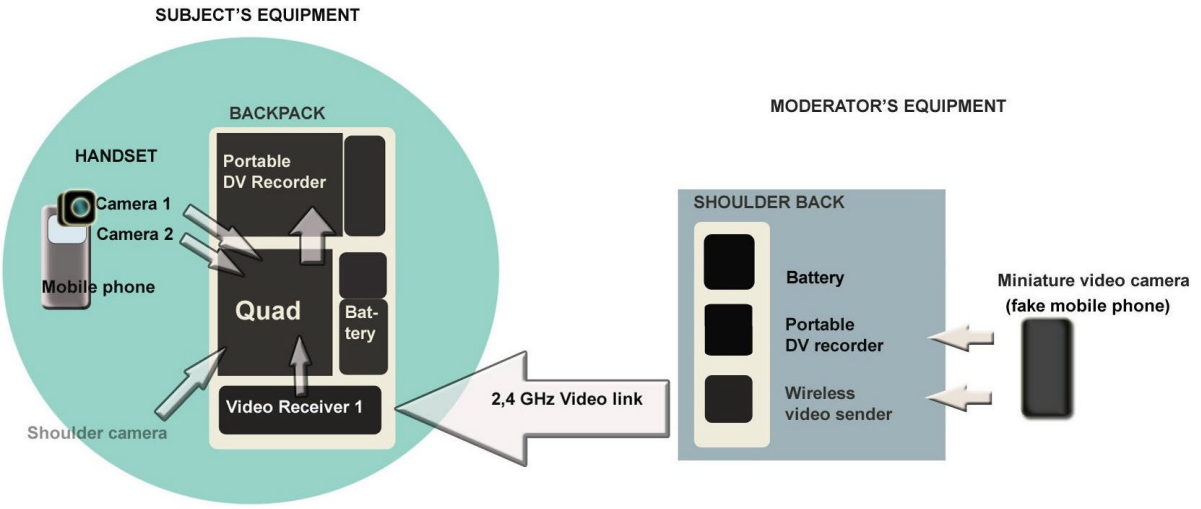

Figure 1. The recording equipment

It was clear from the early stages of the project that it is not possible to follow the user's actions and the environment in detail during the experiment. We needed to videotape the sessions and capture the data afterwards. As the key elements, we used four miniature cameras, Watek WAT-230A. This camera type was selected because at a weight of only 30 grams, the additional weight attached to the mobile phone was 80 grams altogether. The camera model also had a robust design that enabled the use of many different lenses with the element.

Images from the four different cameras were combined into one real time with a digital quad processor Artnix, placed in the subject's backpack. This processor was not as light to carry as it could be, but it provided good image quality with many functionalities and alternatives, such as the date and time displayed on the video. The video stream from the quad processor was recorded with the audio to a Sony mini DV camcorder, carried by the subject.

One lavalier microphone, Vivanco, was used for recording the environment sounds. To protect the microphone from rain, and to minimize the amount of visible wires, it was located on top of the subject's backpack. A preamplifier (Vivanco) was needed to get the signal into the camcorder line-level audio-in.

In addition to the devices, we had to carry a set of batteries. We had two sets of batteries, so we were able to recharge one set while the other set was in use.

3.2 Subject's Apparatus and Materials

The participants carried most of the equipment in a backpack. It contained the microphone, the video camcorder, several batteries, the wireless link receiver, and the quad for building up one video of several video streams (Figure 1).

The test tasks were conducted with a Nokia 6600 mobile phone running a mobile Web browser, Opera by Opera Software. Detailed user actions on the phone were captured with a miniature camera specifically developed for mobile phone user interface research. The miniature camera was attached to the test phone, capturing the phone display and key input. The device was also equipped with a second camera head that was focused up towards subject's face, to see if s/he looks at the phone or the environment (Figure 2).

Figure 2. Nokia 6600 with 2 minicams

A third camera was attached to the backpack shoulder strap, facing the front of participant, in order to record the approximate same view as the participant did (Figure 3).

There were several batteries, as the equipment was designed as a quick and modular construction. The WAT-230 camera elements with 6V operating voltage had their own battery holders: four AA batteries each. The Sony camcorder had batteries of its own standard. A 12V rechargeable lead acid battery pack was needed for operating the quad, audio preamplifier, and the wireless video receiver.

Figure 3. Our portable equipment consisted of 4 miniature cameras,

a microphone, and the recording equipment.

3.3 Moderator's Apparatus and Materials

The moderator was equipped with a miniature camera attached to a fake mobile phone, a camera bag, a camcorder, an A5 sized test script on paper, and a pen.

Moderator's miniature camera was used to record the overall picture of the environment. The video stream was sent wirelessly to the receiver in the subject's backpack. Since we knew that the quality of wireless video is not always comprehensible, we also recorded this view onto videotape carried by the moderator. We have not needed to view these tapes, fortunately. Viewing this tape in sync with the other video material would need special equipment.

Sticker tape on the fake mobile phone helped the moderator to fasten the device on his/her clothes or under the strap of the camera bag for note taking.

In the camera bag, there was a video camcorder that showed and recorded the picture of the moderator's camera. The moderator could check the recorded image by lifting the flap of the bag.

The materials were divided into two parts: guides and forms to be used in the laboratory, and a test script carried on the field. The laboratory guides consisted of a equipment guide explaining the steps needed to prepare the test apparatus for the session, and a pre-test guide to be gone through with the subject before the session.

The test script was A5-sized to make it easy to use on the field. The script contained pictures of places where the tasks should be given to the participant, tasks with the corresponding bookmark number, right answers to the tasks, and approximate task times. In the script, there was also space for note taking and for participant's answers.

3.4 Subjects

The adequate number and type of subjects is highly dependent on the goal of the study, so we are not going to describe the subjects of this specific experiment in detail. The difference between a laboratory study and a mobile context study is that you need to ask the users to be prepared for a session outdoors, perhaps inquire or advise upon their clothing beforehand, and verify that the participants would be physically able to carry the needed equipment. The reliability of the mobile equipment is hardly on the level of fixed one, so the facilitators need to be prepared for cancellations. We invited 32 participants, but managed to execute 27 sessions.

3.5 Tasks

Compared to a laboratory test, designing test tasks for the mobile context is a challenging and time consuming piece of work, because the tasks have to be mapped to the different locations. If the subject is moving while executing the task, we have to either verify that the task will be finished before the next stage, or prepare for cancelling the task or movement when getting there. Many times, the order of tasks has to be counterbalanced to avoid the bias of having a single task type in a certain context. The locations cannot be reordered easily, but the more the order of locations can be mixed, the more reliable results. In our study, we travelled the route to the opposite direction.

If the tasks are different in different locations, we cannot plan having tasks which are location-dependent. This makes it harder to invent natural test cases. Also, tasks including text input should not be used carelessly, since text input is very difficult while moving.

All the above makes it laborious to map the tasks to the locations.

3.6 Procedure

Before meeting the participant, the moderator prepared for the trial by checking the apparatus, rehearsing the overall structure of the session and instructions to be given to the participant at different phases, as well as reviewing instructions on videotaping the trial and on interaction with the participant. In particular, moderators were instructed to read aloud the task instructions from the test script and only follow the participant and videotape, without obstructing participant's view to the environment, where he/she looks at.

Before the actual test, the moderator greeted the participant, committed to paper relevant information about her/him, and read aloud an overall description of the experiment. Next, participants were trained on using the mobile browser and the mobile phone. Training was incremental in nature, starting from simple tasks (e.g., opening the application menu) and ending at two full tasks (e.g., looking from whatis.com at what "ITV" means).

- The task was read aloud to the participant right before starting the execution. Each task was instructed to be performed in one of the "temporal tension" conditions: 1) in the *hurry* condition, the task was to be accomplished as quickly as possible. 2) In the *deadline* condition, the task was to be done within a given time frame, e.g., 4 minutes. The timeframe was enough to perform the task. 3) In the *waiting* condition, the participants were waiting for a metro, for example, and had enough time to carry out one task. After accomplishing the task, the participant reported the answer to the moderator, who then wrote it down on the test script.

In the post-session, the participants were given a form instructing them to recall the locations and answers to each task they performed on the route.

Finally, the participants were thanked for their participation, told the real purpose of the experiment, and asked not to talk about it to the other participants.

3.7 Route Selection

In comparison to laboratory experiments, controlled experiments in the mobile context call for selecting a route for the sessions. This route is partly for purposes of stimulus materials but also partly for procedure as the places have to be visited in a certain order that also determines the order of stimuli.

Our route selection was based on the following criteria:

- *Reliability*. The route had to be reliable and predictable. For example, we could not use rare or infrequent bus connections.
- *Variability and Richness in Visited Places*. The route had to contain enough variation in the mobile places visited.
- *Small variance in duration*. The route had to be designed so that at different times of day it would not be dramatically slower or faster than in the other.
- *Reversibility*. Because we had to minimize the possibility of traverse order biasing our data, the route had to be traversable in the opposite direction as well.

- *Resting Places*. The route had to contain places suitable for resting and for checking that the equipment was functioning ok.
- *Cover from Rain and Cold*. Because people preferred to use the mobile browser without gloves even when it was -15°C, and the equipment was not waterproof, we had to limit the periods spent outdoors.

If a test takes a longer time than expected, it is not trivial to just skip the rest of the tasks and move on to the post-session. It takes time to get to a place where the post-session can take place. This should be taken into account when specifying the route and timings.

4 Analysis

4.1 Coding

Videotapes were converted from DV-tapes to digital format using DIVX3-codec. The test moderator coded the actions and events on to Excel data sheets. This was done by hand by watching the video (Figure 4), pausing the playback after each meaningful event, and recording it on the sheet.

We did not fully follow the categories by Hoyoung [6], because we had a relatively short test session with predefined tasks and a single mobile device meant to be used with one hand. We wanted to focus on the movement of legs, the co-location, and interaction. For these variables, we used a wider value range than Hoyoung, as listed below. In addition to these, we had a category – "temporal tension" – indicating whether Mobile Internet was used in a hurry, within a predefined time frame or at time when no time limit was imposed on Mobile Internet usage.

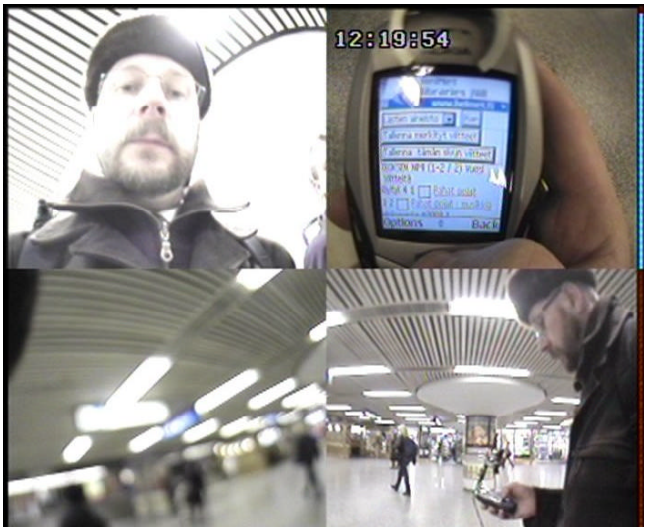

Figure 4. Coding was done by recording the events from one video, consisting of 4 camera views.

The information coded was:

- Time stamp: accuracy of one second
- Task number
- Location: Café / metro platform /…
- Tension: Hurry / Deadline / Waiting

- Leg movement: Normal walk / Decelerated walk / Stand / Sit
- Focus of the user's attention: Phone / Environment
- Interaction: user starts operating the phone / stops it
- Status of the phone: Loading / Scrollable / All content loaded
- Social environment: No people around / Some people around (not moving) / Some people around (moving) / Crowded
- Other

4.2 Data Preparation

As each observation row in the raw data contained often only one change in comparison to the preceding row, most columns were empty. First, empty cells in the data sheets were filled with information from the preceding rows, so that each row included all information describing the situation at that moment.

The second step was to create taxonomy of places and replace place names in the data with the corresponding category name (e.g., "metro escalator at Ruoholahti" and "metro escalator at Rautatientori" were replaced with "metro escalator"). By doing this, we imposed a categorization of places that may or may not be faithful to the individual characteristics of each place.

The third step was to associate participant background information with the data, so that correlations could be calculated between participant characteristics (e.g., age) and dependent variables.

The fourth and final step was to calculate "transformations", by which we mean calculating the frequency of events and their duration during transformation of action or in context to another one. For example, we were interested in how many times attention switches to the environment during page loading, which required calculating a transformation from "page loading starts" to "page loading ends".

5 Discussion

For a significant period of time, it has been understood that laboratory experiments are not fully valid for testing and for studying interfaces intended for mobile use. Ecological validity has been the main objection and several methods have been explored that address this problem. As reviewed in the Introduction however, there are reasons of comprehensiveness, obtrusiveness, and control that limit their ability as a suitable mobile experimentation method.

In this paper, we have proposed quasi-experimentation – conducting comprehensive controlled, and relatively unobtrusive experiments in the wild as the closest match for laboratory experiments. The best benefit gained from using quasi-experimentation is the improvement in our ability to establish causal relationships between contextual events and interaction. There are many threats to the construct and internal validity, but as shown in this paper, there are potential countermeasures. However, as it now stands, quasi-experiment is a laborious method and requires careful planning and vast technological and HCI resources.

To conclude the paper, we first discuss some threats to validity of quasi-experiments and then some of the more practical problems encountered in our work. In doing so, we want to suggest improvements to the method.

5.1 Validity of the Data

In the following, we list distinctive and typical threats in mobile experimentation and discuss how they were addressed in our experiment (full list of threats is presented in [3]):

Threats to *statistical conclusion validity*:

- *Low Experimental Power*. Due to the high level of "noise", experiments in the wild are susceptible to low power. This threat can be addressed 1) by ensuring that each experimental condition receives a substantial amount of data points and 2) by using stronger experimental designs. For the first point, we used 32 participants who all contributed over 1.5h of video data. For the second point, we used a within-subject design where every subject contributes data to all experimental conditions.
- *The Reliability of Treatment Implementation*. When different moderators behave differently on different occasions, error variance will increase and the chance of obtaining true differences will decrease. This problem was addressed by carefully guiding all five moderators on the field with step-by-step instructions.
- *The Reliability of Measures*. Measures of low reliability may not register true changes. This threat can be addressed by employing several people to code the same data with the same instructions and having them check the intercoder reliability with a statistical estimate. Depending on the experiment and the statistical estimate, a practically reasonable reliability is about .80. Acquiring intercoder reliability measures is understandably laborious. Therefore, in our case, in the future, we intend to check only the variables that are influenced mostly by subjective opinions (e.g., "crowdedness" of a street), whereas more "objective" measures (e.g., how long attention stays on the mobile phone) can be left without reliability checking.
- *Random Heterogeneity of Respondents*. Exceptional participants may bias observations in the conditions they take part in. This threat is addressed by using a within-subject design where every participant "biases" each condition equally.

Threats to *construct validity* (cause and effect validity):

- *Mono-Method Bias*. Single method for measuring causes and effects under-represents the construct. In our experiment, we were worried that videotape coding could miss some external events that were actually causing the observed behaviors. To ensure that other possible explanations were not missed, the coding scheme included an "other" column for describing events taking place that did not fit the coding scheme. This is particularly important in quasi-experiments.
- *Experimenter Expectancies*. The data in an experiment may be biased in the direction of the experimenter's expectations. Particularly, experimenters might be susceptible to the influence of mobile contexts and this might be carried over to how they instruct and respond to the participant. Although this threat is almost impossible to eliminate completely (perhaps by replacing moderator by a computer?), we aimed for minimizing its possibility by instructing moderators on how to behave and how talk to the participants uniformly in different conditions.
- *Confounding Levels of Constructs*. When the relationship between an independent variable and a dependent variable is not linear, but only one or two levels of that variable are manipulated, erroneous conclusions about its impact can be easily

drawn. Consequently, in the results analysis, we have to treat levels of variables (e.g., location) as nominal rather than continuous variables.

- *Evaluation Apprehension*. Apprehension about being evaluated may result in attempts by respondents to depict themselves as more competent than is in fact the case. This is true of all experiments, be they conducted in the laboratory or in the wild. However, this threat is particularly imminent in experiments, such as ours, where participants faced a new device and their behavior was videotaped. As in the laboratory tests, we attempted to reduce the threat by highlighting natural use, by minimizing the moderator interference in the usage, and by a relative long usage session of 1.5 hours. However, we did not manage to nullify this threat.
- *Interaction of Setting and Treatment*. In our experiment, an obvious manifestation of this threat was that results obtained during one day might not be obtained on another day. Particularly, some trials were carried out during weekdays and some during weekends. Moreover, they were conducted at different times of the day. Of course, mobile contexts can differ radically between these times. Optimally, all trials should be carried out during the same weekday and at the same time of day. However, due to limitations in experimental technology (one set) and a tight schedule, this was not an option.

5.2 Subject's Apparatus

The subject's apparatus were convenient and handy for carrying around in the city for 1.5 hours. The apparatus thus provided diverse video material about the tests: picture from four angles and audio. The cameras offered sufficient information on what happened on the field, including user-phone interaction, phone status, user's eyes, and the environment (Figure 6).

Environment camera: The test was intentionally planned to collect redundant video material from subject's and moderators environment cameras. The environment video recorded by the moderator was sent wirelessly to the recording videocamera in the subject's backpack and it was often prone to interference from the environment (e.g. in metro escalators). Subject's environment camera, which was fixed on the right strap of the backpack, was not always providing useful data. If the participant happened to be larger or s/he wore a thick winter jacket, the environment camera pointed too up (Figures 4,5). Sometimes the collar of the jacket covered the camera. The user's environment camera fastening should be improved, so that it could be adjusted to the correct vertical angle. Or, if the camera was less noticeable, it could be placed on user's forehead. We did not dare to do this, because we did not want the equipment to make the participant feel uncomfortable.

Figure 5. Miniature cameras recorded the events also when the subject was not using the system

Mobile phone and the attached mini-cameras: The participants used the mobile phone with the mini-camera combination relatively naturally, even though they could not slip it into a pocket due to the cameras. The cameras were small, light, and good enough quality to play a key role in this experiment. They did not stand out too much or interfere with the browsing tasks. Answering an incoming call with the attached cameras would have been impossible, though.

It was not surprising that the phone-camera combination caught some eyes while on the move in the city. In a few cases, a fellow passenger or people in the café asked what was going on and what was the phone-camera combination. The participants politely answered the questions and continued with the tasks. The questions did not seem to bother the users, but of course it affected their attention focus, which we were examining. Naturally, the less noticeable the equipment is, the better.

Audio: Before the test, we expected the shadowed user not to speak a lot, so it would be more important for us to hear the environment, rather than the user. This is why we placed the microphone in an easy position right above the backpack, where it was protected from possible rain and the amount of visible wires could be minimized. When analyzing the test videos, we noticed the importance of recording the user's voice, however. For example, listening to user's comments eased the coder's task of following the test, especially in problem cases, and made it clear what the user's answer was to the test tasks. Unfortunately, the recorded voice was rather quiet due to the location of the microphone. Thus the subject's apparatus could be improved by placing the microphone in front of the user, perhaps to one of the backpack straps.

The weather: The cold winter weather conditions caused some inconvenient problems for the tests. Firstly, the participants had to use the mobile phone outside in -15 degrees Celsius, making phone use somewhat clumsy. Secondly, the duration of the recording equipment batteries was slightly reduced due to the cold weather. Thirdly, the sunny daylight caused reflections on the phone thus making it difficult for the users and the video coders to see the display contents. The use of the phone outside was minimized with careful planning of the test route and altogether it was used outside less than ¼ of the time. The use of the phone could be improved with an automatically changing contrast, depending on the lighting conditions.

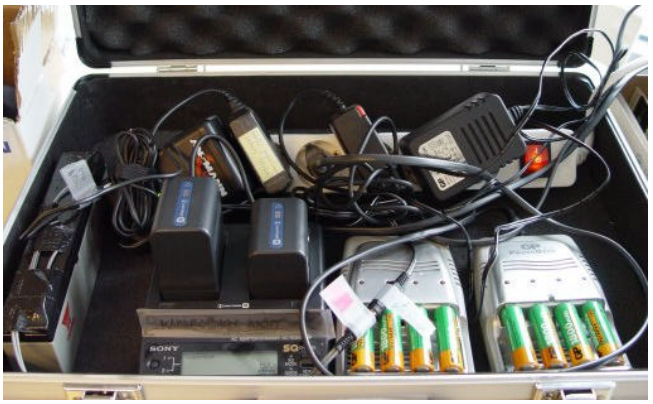

Figure 6. One set of the batteries was being charged while the other set was in use

5.3 Moderator's apparatus and tasks

The camera controlled by the moderator was useful, but there is still much to improve with the camera. First, the technical quality of the wireless video was not always the best possible. Second, the camera's battery connector was too loose and sometimes detached by accident, causing the picture to disappear. Third, a display that shows the environment view being recorded would help the moderator to estimate what s/he is recording. The camera view was shown in the camera bag only and the moderator could not follow the view easily.

When the moderator was taking notes with pen and paper, it was difficult to keep the camera in hand. The plan was to attach the camera to the inner side of the shoulder bag strap with sticker tape for hands free use. Opinions about the usefulness of the sticker tape diverged: some moderators found the sticker tape so practical that they kept the fake phone with the camera behind the strap for long periods, but the other moderators found the sticker tape disturbing.

A small and practical test script was essential; an A4 sized script would have been too big. A hard writing pad, on which the moderator could fasten the handouts, the pen, and the camera, would be useful. A touch screen device might also be suitable for following the task instructions and for note taking; especially if the miniature camera could be attached to it.

One camcorder provided a view to all four cameras on one screen, but this screen was in the backpack carried by the subject. It would be better if the moderator could follow the recorded material during the test, so that s/he could better answer to participant's questions and make a decision to get on with the next task in problem cases. This would also help to recover quickly from technical problems with the video recording. Unfortunately, transferring the views from subject's cameras to the moderator would need either wires between the subject and the moderator or high-quality wireless video transmission.

The content of moderators' recordings were not consistent, although the goals of recording were discussed before the experiment. Some moderators recorded mainly the participant, others more the environment. A question was posed as to what was actually allowed to record, since videotaping the environment might not be allowed everywhere. It is important to discuss these things beforehand and decide what to record in different places.

Despite of the detailed instructions on how to recharge and use the apparatus, the battery recharging process and the equipment turn-on process was complex. There were 6 different batteries in both battery sets, so the moderator had to be careful and avoid mixing the

charged batteries with the empty ones (Figure 6). If there was only one set of batteries, the test schedule would have to be planned to include enough of time for recharging the batteries between the sessions.

To start the recording, the moderator had to connect 6 numbered wires. To save time and to decrease the number of mistakes, it would be good if the power could be switched on and off with one button. We had some contact problems with the connectors, so having a single point in each bag to turn on the power might decrease these problems also.

5.4 Coding the Videotapes

Videotapes were converted from DV tapes to the digital AVI format using DIVX3-codec. This helped us to divide the coding tasks between the moderators and allowed us to work from home. The downside to this digital format was the fact that picture quality was slightly lower than on the original digital videotapes.

The coding phase was the most time consuming part of the whole experiment. One test session took approximately 2 hours, but analyzing one test video took 4 to 6 hours.

Dividing the coding task between the 5 moderators was a great idea for two reasons. First, the moderator was already familiar with his own test and thus could more easily code the video. Second, the coding could be done simultaneously to save calendar time. Having several different people conducting the video analysis caused, however, variance in the interpretations of e.g. events and locations**.** This meant that before starting the data analysis, the data sheets had to be made consistent, as described in the Data Analysis chapter. A strictly formalized data sheet template would minimize the variation in event coding.

We analyzed the videos manually by pausing the playback after each meaningful event and marking it on the data sheet. In indistinct situations, and when there were many different events to be coded, the same part of the video was replayed several times to code it correctly on the data sheets. This could be eased a bit by using automatic logging for the events and for the environmental conditions, as described below.

User's actions and the status of the phone could be recorded easily and accurately if the mobile phone was equipped with a logging application. With time stamps, the log could be combined with the other data. Unfortunately, we used a 3rd party mobile Web browser, Opera, so we could not build a logging system into this software.

User's attention could be observed by using a portable eye tracking equipment during the test, or a suitable image recognition method in the video analysis phase. Both systems are very expensive, not completely reliable, and would require training.

Automatically detecting the environmental conditions might help in documenting the changing variables during the test session. However, having all needed detectors would make the system even more complex, expensive, risky, and likely impossible for female subjects to carry. Also, since the today's detectors are not likely reliable enough, we have to code the environmental changes by hand for the present. It would help, however, if the test moderator could use a separate apparatus, like a palmtop computer, to record these events already on the field.

6 Conclusions

User studies in the mobile context are essential in to understanding human behavior "in the wild" – outside the laboratory environment – and to helping develop systems that fit into this context. In this paper, we described and analyzed a method and apparatus for carrying out a user study to record usage of a mobile phone application, user's eyes, and the environment. The method is called quasi-experimentation [3].

Our portable apparatus allowed us to gather multi-variable data from the different mobile contexts. We have been unable to locate any other studies where data about this number of variables would have been gathered from a user study in the mobile context. Our method is suitable for field experiments where not only device interaction, but also user behavior and the environment, have to be examined.

We managed to root out relevant and scientifically reliable data about user behavior in the mobile context, so the method was successful for our purposes. However, research and development of methods for investigating mobile interaction is necessary for the advancement of mobile computing systems and applications. We need more easily operated portable, light equipment to be able to record the events and environmental circumstances automatically, thus minimizing the manual coding effort. Future work will hopefully be able to provide better and lighter-weight methodology for system developers and other practitioners to evaluate and examine their applications out in the wild.

Acknowledgements

We want to thank Jaakko Aspara, Tero Jantunen, and Sakari Tamminen for helping to conduct this laborious experiment successfully. Nokia Corporation's Multimedia Business Group funded this study. The Academy of Finland supported the second author.

References

1 Cheverst, K., Davies, N., Mitchell, K., Friday, A. Experiences of Developing and Deploying a Context-Aware Tourist Guide: The GUIDE project. Proc. ACM International Conference on Mobile Computing (2000).

2 Consolvo, S. Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Computing, Special Issue on Human Experience, April-June 2003, 24-31.

3 Cook, T.D., and Campbell, D.T. Quasi-experimentation: Design and analysis issues for field settings. Rand McNally (1979).

4 Czerwinski, M., Horvitz, E. and Wilhite, S. A diary study of task switching and interruptions. Proc. CHI 2004, ACM Conference on Human Factors in Computing Systems (2004).

5 Fogarty, J., Hudson, S.E., and Lai, J. (2004). Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. Proc. CHI2004, ACM Conference on Human Factors in Computing Systems (2004) 207-214.

6 Hoyoung K., Jinwoo K., Yeonsoo L., Minhee C., Youngwan C. An Empirical Study of Use Contexts and Usability Problems in Mobile Internet. Proc. 35th Annual Hawaii International Conference on System Sciences (HICSS-35'02) (2002).

7 Iacucci, G., and Isomursu, M. Facilitated and performed "happenings" as resources in experience design of tangible computing. CHI2004 Workshop Cross-dressing and border crossing: exploring experience methods across disciplines (2004).

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

9 Lyons, K. & Starner, T. Mobile Capture for Wearable Computer Evaluation. Proceedings of IEEE International Symposium on Wearable Computing (2001).

10 Oviatt, S. Multimodal System Processing in Mobile Environments. Proc. Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST'2000), 21-30.

11 Tamminen, S., Oulasvirta, A., Toiskallio, K. & Mäkelä, A. Understanding mobile contexts. Special Issue of Journal of Personal and Ubiquitous Computing. Forthcoming in 2004.

12 Woodruff, A., Aoki, P. M., Hurst, A., & Szymanski, M. H. Electronic Guidebooks and Visitor Attention. Proc. International Cultural Heritage Informatics Meeting 2001, 437-454.