Usability takes a hike!

Rigour meets relevance when usability goes mobile

Tomas Lindroth, Stefan Nilsson & Per-Ola Rasmussen

P-O, Stefan, Tomas
@laboratorium.htu.se

Gothenburg University
Department of Informatics

Abstract

Computer power that formerly only was available in offices and homes have now moved out on the roads, seas and beyond. Everything that can be mobile will be, and today only our fantasies are the limit as to what mobile devices can and will do. Mobile devices can be used for anything from taking notes in a business meeting to track down giraffes in a field study in Africa.

When we do traditional usability tests on applications using stationary computers the context is controlled and not especially relevant. The computers in the labs are more or less in the same context as when they are used in offices and homes. But for mobile devices, testing might make the result irrelevant since it fails to take the context of its use into consideration. The purpose of this thesis is to evaluate the usability testing methods and theories from a mobile perspective. This is to find out if and where the conventional usability methods fail and what they fail to detect when applied to mobile devices. How can the usability methods of today be extended to facilitate the testing of mobile devices in its right context?

We have done empirical tests of usability methods in usability laboratories and conducted expert interviews with researchers from the mobile as well as the usability field. Together with literature studies and informal interviews we analyze and discuss around rigour vs. relevance in laboratory and mobile settings. We used triangulation on the usability methods we tested and combined these results with the results from the expert interviews. First of all we found that there is indeed a need for a way to conduct mobile usability testing. The conventional usability tests take little or no consideration to the context of its use. All it measures is how good the gadget is in an office-like environment like in for example a usability lab. We propose a new tactic for usability test mobile gadgets. The tactic consists of conventional usability methods combined with an ethnographical study of the use of the gadget in the real world and a role-playing part where made up tasks are conducted in real life. This is to combine both rigour and relevance in the testing and introduce contextual aspects.

Index

1 Introduction

Computer and communication technology has evolved from stationary to mobile. From controlled context to uncontrolled context, from the predictable to the unpredictable. Small screens have replaced big screens. Pen has replaced keyboard. These are all big changes, changes that alter the way we use the computer in a fundamental way. During a long period of time we have seen stationary and static offices turn into mobile and diversified workplaces. Computers and cellular phones have evolved in such way that they offer the same possibilities to perform advanced tasks in the field as before had to be done in an office. Every day the list of performable tasks possible gets longer concerning mobile devices. We no longer have to sit in an office or at home to perform task of the everyday life. We have come to a point where mobile devices have the same power as stationary computer in the office. In theory we can work anywhere we want, anytime.

The mobile device is seen as remote control for business and pleasure where you can buy, sell, control and supervise any gadget or situation. Without designers with the proper knowledge about HCI (Human Computer Interaction), Information- and User centered design there is high risk of usability flaws sneaking in to the design. With mobile solutions based on devices such as PDAs and cell phones the design of the gadgets and their interfaces are crucial factors for success. If designing for the web is hard with different browsers, screen sizes etc, try designing an interface on a screen with the size of half your credit card that might be used on the run in a dark alley with the rain pouring down. It is a possible scenario, mobile really means mobile, and it really means anywhere, on the bus, at the beach or in a storm. Testing of a new website is a must with different browsers, connections and users. But testing in front of a computer in a controlled environment is one thing, testing for mobility another. Usability testing in a laboratory with controlled situations and tasks works for applications used in stationary solutions. In the lab there is possibilities for video recordings with sound, screen captures, observers and controlled tasks. As expressed by Johnson (1998), this works fine with solutions where the context and environment is of second interest. Now, think of usability testing of a mobile solution where context is a factor.

"It is cold and snowing and you do not know from where your bus leaves in 5 minutes. You pick up your WAP phone to check: The mobile user run to catch her bus, after her run three researchers with cameras and microphones."

The idea for the subject of this thesis has developed over a long time. All of the authors have over a period of a couple of years been exposed to an environment where mobile gadgets have been commonplace and used daily. It has been a part of our work to look at these devices from a critical point of view. In this process, we have often found situations where we wanted to use the gadget and discovered how hard or even impossible it was to use them in the specific situation. It could be a lack of light or too much light, rain or sunshine, high or low temperature.

This got us all to think of why this happened so often. There are undoubtedly several answers to this question. Some of the weaknesses in the gadget we found could be traced back to a lack of available technology – it simply did not exist at the time of the design of the gadget. Another

could be that the designers wanted to make the gizmo as small and light as possible. Maybe they simply had not designed it for the context we wanted to use it in.

1.1 Purpose

We will in this thesis explore current usability methods that are used today to test usability aspects of stationary computers. We will also conduct interviews with experts in the fields of mobility. This is done to explore the strengths and weaknesses of the usability methods used today and to aid us in the proposed design of a new way of testing the usability of mobile gadgets in the future, where context is of great importance.

1.2 Question at issue

This master thesis explores available usability methods. This is to see if they can be used in a mobile context.

Our question at issue is:

-Is there any use for new or modified methods when it comes to evaluating mobile gadgets that are used in their right context, and how would such a method look like?

1.3 Demarcation

We have on purpose avoided elaborating on the subject of mobility. This is in itself a vast area of research. Since the focus of this thesis is not mobility of itself, but rather context of use, we have tried to only briefly discuss mobility.

1.4 Terminology

1.4.1 Context

There are several different definitions of what context is. It is commonly used in the meaning of location. But this does not suit us, as we regard context as being a much broader concept. We mean that context is information that in some way affects both the user and the use of the gadgets.

1.4.2 Mobility

"Mobility is one of those words that are virtually impossible to define in a meaningful way. You either come up with a definition that excludes obvious instances, or your definition is to too vague; it fails to shed light on important aspects. At the same time we all have a feeling of what it

means; the newsboy and the travelling salesman are mobile, the secretary and the cook are not. Thus, we can conceive typical situations in which people are mobile and when they are not." (Kristoffersen & Ljungberg, 1999)

As stated by this quote, mobility is not as easy to define as it first appears. It is a vast area of research, and it is not in the theme of this thesis for us to dig deeper into this field. For the purpose of this thesis, when we talk about mobility, we mean the use of mobile gadgets in different situations, not necessarily in the field.

1.4.3 PDA

PDA is short for Personal Digital Assistant, i.e. a handheld computer. A PDA can use a number of different hardware and software configurations.

2 Method

2.1 Scientific approach

The purpose with this master thesis is to explore the field of available usability methods. To se if they can be used in a mobile context. We realized quite fast that we did not know much about this field. Therefore we decided to study the subject in the broadest way possible. The first thing we did was to set up a list of people that we knew had hands on experience regarding this field. We phoned them and shared our thoughts about this subject and our approach to it. This was to se that we were on the right track and to get some sort of confirmation that our approach was scientifically interesting and valuable to the usability field. According to Järvinen (1999) it is of great importance that that the researchers are aware of how, and to what extent, the study is contributing to the scientific field.

We decided to use several different methods to be able to find answers on our question at issue and to be able to achieve our purpose with this master thesis. In our case we used a multi dimensional method, Method Triangulation (Repstad, 1988), which is a kind of hybrid model. We have chosen to combine qualitative interviews, methodology studies and case studies.

According to (Merriam, 1994) there are some different types of ambition levels that you can use in your work. These are explorative (changes), descriptive (objective), comparative, explanatory (understanding) and normative (own conclusions). These can be combined in different ways and in our study we have chosen to be both descriptive and explanatory in the main part of this thesis.

The latter part is written in a more normative way. We think that there is a lot of information lost on the way if we do not share, with the reader, our experience that we have gained during this study. Comments are necessary so that the reader will not get lost among all details (Merriam, 1994). This makes it easier for the novice to understand the thesis and our findings.

2.2 Course of action

The methods used in this master thesis are mainly of qualitative character. We have done interviews, case studies and "on spot" observations. Along with this we have done a lot of literature studies concerning the usability field. Books, scientific articles and websites have been our main information sources.

The first small steps toward our choice of subject for this master thesis were taken in the spring of year 2000. We were a couple of laboratory assistants and scientists that were visiting Aalborg University and a company named Mindpass. At the university we visited their usability laboratory and other computer related departments, for example their "cave". One of the researchers at Aalborg University also worked at Mindpass, Dr. Lars Bo Eriksen. He told us that they were developing a search engine that they would like to test on mobile devices. The question

"How do one perform mobile usability tests" was born.

The next step toward this subject was a meeting during the IRIS-23 conference that took place in August. We had a discussion together with Eriksen and agreed upon a schedule and a research plan. We formulated our goal with the research and planned the staying in Aalborg. We decided to visit Aalborg for four to eight weeks and during that time we would conduct case studies in the usability laboratory and do qualitative interviews.

The time between the conference and our first day in Aalborg was spent on literature studies. We contacted teachers and researchers with knowledge about the usability and mobility area to be able to put a literature list together.

We moved to Aalborg in September. We started to examine the usability laboratory to get used to with the equipment. We performed a couple of tasks on our own to find out what sort of possibilities and limits the equipment had. These tasks provided us with valuable information about how to conduct and plan our usability tests that we were going to do.
In parallel to this we started to do a big methodology study concerning all available usability methods. We also continued our literature study and started to take contact with people in the usability and mobility field to see if some of them could support our work by giving interviews.

All along the work with this master thesis we have had some informal interviews with people from the usability field, mobility field and with ordinary people who use mobile gadgets. Some of these meetings have been recorded. We have asked people in our surroundings how they experience using mobile devices on the run in contradiction to be sitting down and so on. We have asked them what they use the devices for and what they think of them. These reflections have proven very valuable for us during our work and in our planning of tests.

The tests conducted during this time are explained later on in the section "Empirical Study and Analysis" of this thesis.

On basis of drawn conclusions, from the tests and informal interviews, we formulated our questions for the interviews that we were going to do. We decided to send out the questions by email. All participating persons were so geographically diversified so we did not have the possibility to meet them in person, mainly because of lack of money. Beside this, some of them were abroad during the time when we were ready to conduct such interviews. The questions were few and quite simple to answer in a few lines, therefore we decided not to spend time, money and effort on telephone interviews.

2.2.1 Method Triangulation

As mentioned earlier we have used Method Triangulation. In mathematical term, triangulation is a method of determining distance and position by measuring the distance between two fixed points and then measuring the angle from each of these to a third point. By analogy, triangular techniques in the social sciences attempt to map out, or explain more fully, the richness and complexity of human behavior by studying it from more than one standpoint and, in so doing, by making use of both quantitative and qualitative data (Cohen & Manion, 1994). It is a concept-applied triangulation in educational research as a means of countering the selective bias of a

single view (Robinson, 1992).

Triangulation is the use of different research methods or sources of data to examine the same problem. If the same conclusions can be reached using different methods or sources then no peculiarity of method or source has produced the conclusions and one's confidence in their validity increases. (Lwin, CPD, 1997)

Among the different types of 'triangulation', the following are the some examples of the principal types of triangulation used in research. Data triangulation refers to the collection of varied data on the same phenomena, e.g. from different participants, different phases of fieldwork.
Investigator triangulation, similarly involves collection of data by more than one researcher (preferably through adoption of different roles in the field) and method triangulation involves the collection of data by different methods, which entail different threats to validity. (Lwin, CPD, 1997)

Robinson (1992) also suggests that multiple methods or triangulation have been used in educational research when:

1. A more holistic picture is wanted;

2. Complex phenomena are being investigated;

3. A controversial topic is being investigated;

4. Validity and minimization of bias is sought;

5. Established single-method approaches have proved arid, too narrowly selective and uninformative.

For the purpose of applicability, external validity and minimization of selective bias of a single view, we have used three different methods.

2.2.2 Case study

Case studies imply that one study and investigate a small quantity of objects (for example students, companies) in a lot of respects (Ericsson, Wiedersheim-Paul 1997), in our case how a small number of people succeed to perform some different tasks on a PDA. Case studies is a method used when one want to get close to a specific research area/problem area Merriam (1994), partly through observations in natural environment and partly thru qualitative interviews to be able to catch subjective factors. According to Wallén (1996) the main benefits of direct observations is that one study the persons in their natural environment and that one can get profound knowledge about the course of events. The purpose with a case study is to broaden the area of knowledge and to further develop methods and concepts (Wallén, 1996, Lindberg 1999). There are a lot of ways to perform case studies. They can take part in a laboratory as strictly controlled experiment or in their natural environment. We have chosen to perform them in both types of environment.

2.2.3 Methodology studies

It is of great importance that you study all available methods, when you as a researcher are entering a new scientific area and are trying to develop new methods or evolve old ones. We did a thorough methodology study and came up with twenty-two different methods that were applicable in our case. We analyzed them and tried to sort out methods that would give us a good picture of what the different types of methods did/did not measure in different contexts.

2.2.4 Formal and informal interviews

During our study we have had several informal interviews with users and experts in our case study. We have deliberately chosen to keep them as ordinary conversations even though they have been interviews. By doing so we have avoided steering the conversations in the slightest way possible (Holme, Solvang, 1991). We have only set up the thematic frame for the conversations.

The formal interviews that we have performed have been structured and sent out by e-mail. The problem with interviews like these is that different people can interpret the material in different ways. (Galtung, 1967) We think that this is not the case in our study since they have been sent to people that are experts in the area of mobility and usability. The questions have been of such character that they have only given their own personal thoughts/opinions about the questions at issue.

Quotes from the interviews have, if needed, been transcribed and/or translated into English. In this process we tried to stay as close as possible to the original meaning of the statement.

2.2.5 Collecting the material

All research should be based upon, or consider, earlier studies done in the same field of interest. By taking part of earlier research material one can get help with how to define conceptions, plan the research and how to interpret results (Merriam, 1994). If you do not take part of earlier researches and theories, there is a risk that you might repeat others mistake or in worst case produce a copy of an already existing research.

We have been collecting material for this research since August. We have read several scientific papers, dissertations, books and interviews. These have been collected from Internet, libraries, article libraries, scientific conferences and meetings.

2.2.6 Treating the information

We did a literature "walk through" after the collection. We had a very large collection of material and therefore we had to some thinning. We choose to set up some guidelines for this.

- When was the material produced?

- Is it published or is it a working draft?
- Who is the writer and what connection to the area has s/he got? Ex. Scientist, student and so on.
- Type of study? Ex empirical case study, theories, laboratory experiment and so on.
- Quality of the source (structure and underlying thoughts).

2.2.7 Ethic

A basic question in all research is to what extent a research is conducted in an ethical way (Merriam, 1994). Respect for fellow creature is a basic point of departure when it comes to research. This means that we have to protect physical and psychological integrity for those people who contribute with information. The obligation to observe silence must be followed strictly and we must assure that no one can find out who the respondent is. The respondents should not be manipulated to participate on false background. They should decide themselves if they want to participate (Holme, Solvang, 1997).

Our study is not of sensible character but we have chosen to consider the ethical aspects as much as possible. The users and persons we have interviewed have been informed about how this study would affect them. No one has been forced to participate. All information and results from the case study have been treated confidential. Quotes from the interviews have all been used with the explicit permission from the interviewees.

2.3 Quality

The scientist is the primary instrument in a study of qualitative character when it comes to collecting and analysis of information. (Merriam, 1994) We are aware of the facts that we are human instruments and because we are human we can do mistakes and let personal value and attitudes affect the study. A human instrument is, according to (Merriam, 1994), as unreliable as any other instrument. We conducted interviews with people from both side of the field (usability and mobility) and set up a case study to prevent this from happening. You could say that the interviewed people formed our reference group to the case study, literature and methodology study.

2.3.1 Reliability and validity

The outmost purpose when you, as researchers, are trying to generate theories is the discovery of new theories, either in form of explanation or understanding. (Johansson-Lindfors, 1993) The term validity is not used when you perform a study like this since the discovery of notions and theories, explanations or understandings are the finishing end. (Validity is about measuring what you intend to measure) Neither reliability as a notion is used in this kind of study. Due to the fact that other researchers cannot perform the qualitative methods we have used in this study, in the same exact way again (example: the usability laboratory that we used is going to be reconstructed). The possibility to generalize our theories in other situations than the one we have used is therefore low. (Johansson-Lindfors, 1993; Järvinen, 1999) In our case we talk about applicability, i.e. the possibility to apply our theory, explanation or comprehension in similar situations (mobile usability testing) (Johansson-Lindfors, 1993). This is to a certain extent

similar to external validity (Järvinen, 1999).

2.3.2 Discussion around the chosen method

It is difficult to adopt and understand a brand new scientific field. None of the authors had ever worked with usability testing when we decided to do this master thesis. To be able to achieve enough knowledge about this field in such short time we realized that we had to use several different methods. Both the formal and informal interviews gave us a very good ground to start out from and we saved a lot of time since the interviewed persons guided us in the right direction from the beginning. By doing a methodology study on available usability methods we learned a lot about how to conduct test, what traps you could fall into, what you can/can not measure in different situations and so on. The case study gave us on hand experience about how to conduct test, how to use a usability laboratory, what problems that can occur, how the lack of context affect the test situation and so on. We think that this was the best way to conduct a study like this.

A legitimate question at this time is if we could have done this study in a different way. The answer is off course, yes, but to what cost regarding time, money and knowledge? In the beginning of this study we thought of some different way of how to conduct our case studies. We talked about the possibilities to do usability test with reporters and journalists but we realized quit soon that we did not have the right equipment for a study like that. We also talked about the possibilities to spy on people when they are using their mobile devices (Weilenmann, Larsson, 2000). Most methods that we could think of were not possible to perform, mostly because of the fact that we could not watch over the users in a satisfactory way.

3 Available methods (current usability methods)

In this chapter we present all of the applicable methods that we found. They constitute the ground for our methodology study when it came to deciding what sort of usability tests we were going to do.

They belong to three different areas:
- Inspection and evaluation
- Testing
- Inquiry

3.1 Inspection and Evaluation

Method	Purpose	Characteristics
1) Heuristic Evaluation (Nielsen & Mack, 1994)	Identify usability problems early in the design phase. Guidelines vs. design. You can provide the experts with paper mockups, or even just design specifications, and still get a good amount of usability problems discovered before actual work begins.	The evaluator uses sets of guidelines (i.e. heuristics) and compares those with the interface. The heuristics form a checklist that the evaluator uses during his/her work. With heuristic evaluation, it is possible to identify many usability problems and it is possible to evaluate early on in the design phase. It can also be useful when evaluating the style (i.e. look and feel) of the interface. The heuristics are not "optimized" for identification of usability problems concerning efficiency in daily use.
2) Cognitive Walkthrough (Rowley & Rhoades, 1992), (Spencer, 2000), (Wharton et. al., 1994)	Motivating how or why a person would react in a certain situation. Based on assumption about the users background, knowledge and goal. Great for early stages of development because they can be performed using just a system specification as a basis.	With this method an evaluator examines each action in a solution path and tries to tell a credible story describing why the expected user would choose a certain action. The story is based on assumption about the users background, knowledge and goals, and on understanding the problem solving process that enables a user to guess the correct action. Cognitive walkthrough is an inspection method that focuses on evaluating a design for ease of learning, particularly by exploration. It is more difficult to evaluate efficiency in daily use. Problems concerning the content of the interface are rarely identified, due to the evaluator's limited domain background.

3) Formal Usability Inspection (Kahn & Prail, 1994) (Freedman & Weinberg, 1990), (Gilb et. al., 1993), (Wheeler, 1996)	A way to detect errors in the code that the design relies on and documentation defects. The inspector performs tasks and reports any found errors and the lines of code causing the problem. The technique is design to reduce the time required to discover defects in a tight product cycle. Great for early stages since the inspector can work with merely a specification or paper mockups.	Takes the software inspection methodology and adapts it to usability evaluation. Software inspections, more commonly known as code inspections, started at IBM as a way to formalize the discovery and recording of software problems ("defects" in quality jargon, "bugs" in the vernacular).
4) Pluralistic Walkthrough (Bias, 1991)	Looks into how user's react in different situations. Includes user's, developer and usability experts. Best used in the early stages of development, as the feedback garnered from pluralistic walkthrough sessions is often in the form of user preferences and opinions.	This is a test that can be performed early in the design process. Representatives from the three categories meet and discuss usability problems that are associated with the dialogue elements in different scenario steps. Pluralistic walkthrough is an effective method in evaluating the learn ability of a user interface.
5) Feature Inspection (Nielsen & Mack, 1994)	Find out if the feature of a product meets the users need and demanding. Best used in the middle stages of development. At this point, the functions of the product and the features that the users will use to produce their desired output are known.	Features of a product is listed in the sequence they would normally be performed to perform a task. For example to be able to send a mail you will have to push the button New Mail, insert recipient, subject, text and then you have to push the send button. Each set of features used to produce the required output is analyzed for its availability, understandability and general usefulness.

6) Consistency Inspections (Wixon, et. al., 1994), (Nielsen, 1995)	Looks for consistency across multiple products from the same development effort. Best used in the early stages of development, when the initial development work has not progressed to the point where products that require extensive changes to ensure consistency will not require total overhauls.	A good example of products that can be examined is the whole range of cellular phones from a particular producer, for example from Ericsson, and to see if the user interface and I/O is consistent. Consistency inspections begin with a usability professional analyzing the interfaces to all of the products and notice the various ways that each product implements a particular user interaction or function. An evaluation team then meets, and using the usability analysis as a basis, negotiates and decides on the one golden implementation for the usability attributes of each product.
7) Standards Inspection (Wixon, et. al., 1994), (Nielsen, 1995)	Standards Inspection ensures compliance with industry standards. Best used in the middle stages of development, as the actual design is being developed with the given standard in mind.	When performing such inspection usability professional with extensive knowledge of the standard analyses the elements of the product for their use of the industry standard. The professional should be a member of the country's standard organization and acquainted with the certain area.
8) Guideline Checklist (Wixon, et. al., 1994), (Nielsen, 1995)	Guidelines and checklists help ensure that usability will be considered in a design. Usually, checklists are used in conjunction with a usability inspection method. The checklist gives the inspectors a basis by which to compare the product.	There are a lot of guidelines available and they can be used in the way they are published, although you may want to tailor the guidelines to suit the exact issues faced by your product's user. The list should be short because you will use it against a lot of parts in the interface.

3.2 Testing

Method	Purpose	Characteristics
9) Thinking Aloud (Lewis, 1982), (Dumas & Redish, 1993), (Lindgaard, 1994), (Nielsen, 1994), (Rubin, 1994)	Lets the evaluator understand how the user views the system. The method can be used in any stage of development. Gives a lot of qualitative feedback during testing.	The users verbalise their thoughts while using the system. It is especially useful when applied by the designer of the interface since direct feedback from the users on the design can be obtained. Drawback with this method include that it is not very natural for users to think aloud. It is also hard for skilled users to verbalise their decision process since they execute part of their work automatically.
10) Co-Discovery Method (Dumas & Redish, 1993), (Lindgaard, 1994, (Rubin, 1994)	Idealistic for evaluating groupware programs, CSCW products and other products designed to be used by workers in team environments. Can be used during any phase of development.	The users verbalise their thoughts while using the system. It is especially useful when applied by the designer of the interface since direct feedback from the users on the design can be obtained. Co-discovery is a type of usability testing where two participants attempt to perform tasks together while being observed. The advantage of this method over the thinking aloud protocol is two-fold: in the workplace, most people have someone else available for help the interaction between the two participants can bring out more insights than a single participant vocalizing his or her thoughts.
11) Question-asking Protocol (Dumas & Redish, 1993), (Lindgaard, 1994, (Rubin, 1994)	Lets the evaluator understand how the user views the system. The method can be used in any stage of development. Gives a lot of qualitative feedback during testing.	The users verbalise their thoughts while using the system. The most significant difference from the Talk Out Loud method is that the designers are asking questions during the test. The evaluators ability, or lack of, to answer the questions can help the designers to see what parts of the product interface were obvious, and which were obtuse. Any staff working with usability can perform it but it is especially useful when applied by the designer of the interface since direct feedback from the users on the design can be obtained.

12) Performance Measurement (Nielsen, 1993), (Dumas & Redish, 1993), (Lindgaard, 1994, (Rubin, 1994)	Measures whether a usability goal is reached or not e.g. a kind of bench marketing. Should be used in initial stages of design to provide benchmarks for the design process. It is also used during the design cycle to measure the work done thus far against those benchmarks.	User performance is usually measured by having a group of test users perform a pre-defined set of tasks while collecting data on errors and time. The test is usually carried out in a laboratory. Benchmarks are usually devised during initial usability testing, either of a previous release, or of a competitor product. The data collected must be in, or be translated to, quantitative variables.
13) Wizard of Oz (Green & Wei-Haas, 1985), (Boreczky et. al., 1990)	Simulate a full-scale application/device even though it is only a prototype. Fool the user.	An example describes this method best. Assume that you have an idea about a device or an application that you want to evaluate before the development starts. In that case you would create a mock-up device or an application sketch that would act as a real thing. To make it work you have assisting personnel that perform the tasks that the user wants to do. The user does not know that it is a fake. (Example: to simulate a vice recognition software, i.e. word processor, you can have a typist that is writing every word that the user is saying)

3.3 Inquiry

Method	Purpose	Characteristics
14) Contextual Inquiry (Holtzblatt & Beyer, 1993), (Holtzblatt & Jones, 1993), (Beyer & Holtzblatt, 1995), (Beyer & Holtzblatt, 1997),	Contextual inquiry is used to get a broad knowledge about the environment that you are producing the program or device for. This technique is best used in the early stages of development, since a lot of the information you will get is subjective--how people feel about their jobs, how work or information flows through the organization, etc.	Contextual inquiry is basically a structured field interviewing method, based on a few core principles that differentiate this method from plain, journalistic interviewing. Contextual inquiry is more a discovery process than an evaluative process; more like learning than testing. Quit similar to an ethnographic study. Contextual inquiry is one of the best methods to use when you really need to understand the user's work context. Many times, the environment in which people work really influences how people use a product. It sounds like a cliché, but there really are people who print out their email and mark it up with comments before replying.

| 15) Ethnographic Study/Field Observation (Hammersley & Atkinson, 1995), (Wixon & Ramey, 1996) | Ethnographic Study is used to get a broad knowledge about the environment that you are studying. This technique is best used when you are studying complex situations where ordinary methods would miss to detect important details, for example "unspoken acting" i.e. tacit knowledge. A lot of the information you will get is subjective--how people feel about their jobs, how work or information flows through the organization, etc. | Ethnography is a social science research method. It relies heavily on up-close, experience-near participation (not just observing) by ethnographic researchers, often working in multidisciplinary teams. It usually includes intensive language and culture learning, intensive study of a single field or arena, and a blend of historical, observational, and interview methods. Ethnographic methods can give shape to new constructs, new variables, for further empirical testing in the field or through so-called traditional, quantitative social science methods.

The roots of ethnography are in anthropology and sociology but present-day practitioners do ethnography in organizations and communities of all kinds. Ethnographers study schooling, public health, rural and urban development, consumers and consumer goods--any human arena. While particularly suited to exploratory research, ethnography draws on a wide range of both qualitative and quantitative methodologies, moving from "learning" to "testing" while research problems, perspectives, and theories emerge and shift.

Ethnographic methods are a means of tapping local points of view, a means of identifying significant categories of human experience up close and personal. Ethnography enhances and widens top-down views and enriches the inquiry process, taps both bottom-up insights and perspectives of powerful actors "at the top," and generates new analytic insights by engaging in interactive, team exploration of often subtle arenas of human difference and similarity. | |

16) Interviews and Focus groups (Greenbaum, 1997), (Nielsen, 1997), (Templeton, 1994)	This technique can be used at any stage of development, depending on the questions that are asked. Interviews and focus groups are often held at very early stages of development thou, when the product requirements are still not firm. Focus groups are then held to extract user requirements prior to initial design.	This is a data collecting technique where about 6 to 9 users are brought together to discuss issues relating to the system. A human factors engineer play the role of a moderator, who needs to prepare the list of issues to be discussed beforehand and seek to gather the needed information from the discussion. This can capture spontaneous user reactions and ideas that evolve in the dynamic group process.
17) Customer Research Groups (Lynch & Palmiter, 2000)	Customer Research Groups is an effective alternative to focus groups with the same purpose. (Se above)	Groups of 12 to 15 users are invited to come in at the same time. Instead of having a single facilitator, there is a facilitator for each user. The idea is to get multiple one on one discussions rather than a group opinion. The room is typically divided into four or five different exercises. Each user is given a particular amount of time to participate in each exercise before they move on to the next one. An example of an exercise is a card sort of features, i.e. the user prioritizes the features and explains why they ordered them as they did. There are lots of other exercises that can be used depending upon the type of data desired. The method works well because of the large amount of data collected and the involvement of the entire design team.
18) Questionnaires (Nielsen, 1993)	Gives the evaluator information about subjective satisfaction and possible anxieties.	Questionnaires are especially useful for issues concerning user's subjective satisfaction and possible anxieties. Though, it is difficult to get objective results when using questionnaires since the user's answers are based on what they think they do, not on what they actually do.

19) Journaled Session (Nielsen, 1993)	Journaled sessions bridges usability inquiry, where you ask people about their experiences with a product, and usability testing, where you observe people experiencing the product's user interface. This technique is best used in the early stages of development, probably even pre-development, where the information you are attempting to gather is more preferential than empirical.	Journaled sessions are often used as a remote inquiry method for software user interface evaluation. A disk is distributed to a number of test subjects containing a prototype of the software product, as well as additional code to capture (or journalize) the subject's actions when using the prototype. Users perform several tasks with the prototype, much as in formal usability tests, and their actions are captured with the journalizing software. Upon completion of the series of tasks, the users return the disks to you for you to evaluate.	
20) Incident Diaries or Self-Reporting Logs (Nielsen, 1993)	Finds out what kind of problems a user has had during a period of time or what they have used the system/device for.	Users can be asked to keep diaries of their interactions with the system. Typically they record when a problem occurs. What happened, when and how they fixed/came around it. Scales can be used to set a value on the error. Diaries show how often errors occurs. Diaries can also be used to gather information about the use of a system or how a device is used in an every day situation. A possible scenario is to ask the user to write down what they have used the system/device for, every day during a month. It is a cheap method for gaining information about problems and can be used for a long-term data collection. It does require a level of trust and a level of cooperation from the user. It is not something that should be undertaken lightly since it does place a considerable burden on the user.	

21) The Valuation Method	Finds out how important a feature is to a user.	A user is asked to use a feature on a system or on a device and then to rate how much more they would pay for the feature if it performs in a particular way that the user deems important. This method is useful during requirements gathering to find out what users want and how they rate those wants. However, it should not be taken to seriously as a measure of real prices that people are willing to pay. It is more likely that it delivers a list of priorities and should be treated as such. It should also be remembered that people could rate functionality differently according to what they are doing at the time.	
22) Logging use (Nielsen, 1993)	Gathers information about use and problems without the user knowing about it.	A logging function is implemented in the system or in a device. The log contains a lot of information about how a system/device is used. The biggest problem is to sort out information that is relevant because of the huge amount of data recorded. Another problem is that the logging system does not know in what situation the system/device was used.	

4 View of the usability field

4.1 Rigour vs. Relevance

According to Mason (1988) there exist two primary attributes of knowledge producing activities in controlled experiments. He identifies them as: tightness of control and richness of reality. These attributes are taken generally to be in opposition to one another at the same level of knowledge, called the iso-epistemic curve. Hence, researchers must ultimately make a trade-off between them.

The larger the number of factors that is under control in an experiment, the more scientific rigour is emphasized. The more natural like the experimental setting is, the more relevant and applicable the results will be. (Järvinen, 1999) (see figure 4.1)

Rigour Versus Relevance

Relevance

Relevance= Richness of worldly realism

Knowledge domain defined by theory and conceptual variables

B

Iso-epistemic curve

A

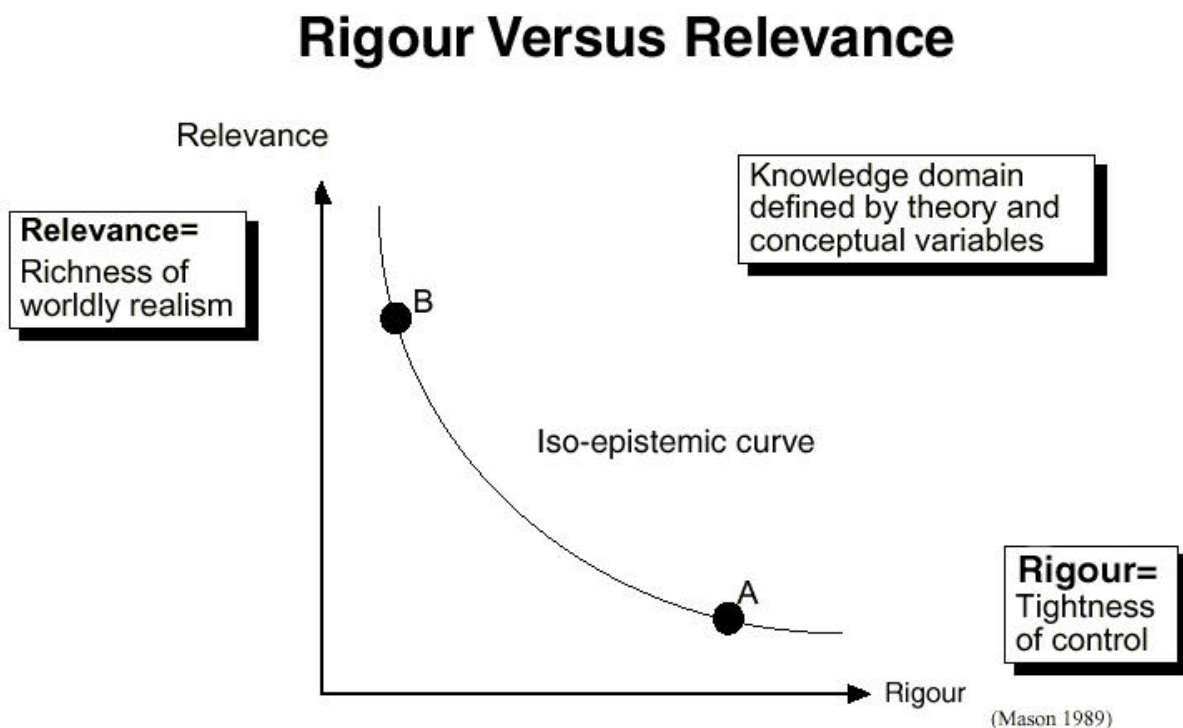**Rigour=** Tightness of control

Rigour

(Mason 1989)

Figure 4.1

4.2 User Centred Design

To make usable products, tools and applications there are several methods and theories that help the designer to reach her goal. They all focus on the user, her needs and requirements. User

Centered Design (UCD) is a process that puts the user and her tasks in focus from the very beginning. The alternative to UCD has been a problem based approach where the user has to fit the solution, UCD demands deep understanding of the users needs and goals (Shneiderman, 1998). A big part of UCD is the iterative design cycle where a solution is designed, tested and modified repeatedly like a spiral (Rubins, 1994). The focus in this article is not UCD in particular but a part of it is the testing and evaluation of the software, device or other product that measures the usability of the same.

UCD means many things and goes by different names, but they are all names of the same concept, design that focus on the user (Rubins, 1994)(Nielsen, 1993). According to Nielsen usability is part of UCD and stands for the evaluation, change and improvement of a system, product or gadget Usability are not UCD, but one of the techniques to secure a user centered design.

4.3 Usability

Usability is the process of testing with a handful of techniques to gain learnability, efficiency, memorability, less errors and satisfaction (Nielsen 1993). These five attributes are the basics of usability engineering according to Nielsen (1993). There are others with their own definition of attributes like Rubins (1994) for instance. He outlines four similar attributes, usefulness, effectiveness, learnability and attitude (Booth, 1989 in Rubins, 1994). These are similar to Nielsens but with a slightly different definition. Without further discussion we choose Nielsens definition because it is the most widely known of these two (Olsson, 2000).

? Learnability

   It should be easy to learn a new system so the user can start working quickly.

? Efficiency

   A system should be efficient to use so the user achieves high productivity.

? Memorability

   A casual user should not need to re-learn between times, the system needs to be logical.

? Errors

   The system should stop the user from doing errors and if the user makes errors she should easily be able to recover.

? Satisfaction

   Using the system should be pleasant. The user should want to return and like to use the system.

Here we use these five attributes as our definition of usability engineering. Any method or theory that supports and enhance one of these attributes would fit into the description of Usability Engineering. These attributes and theories are meant to support rigour.

In the method part above a vast amount of different methods was lined up with a short description of how they work. These methods are merely tools to measure the five attributes above. The product of the different tests is for some methods lists of errors made and for other methods it is videotapes from where you can collect user statements and interesting observations.

4.4 Drifting

We use scissors for cutting, chairs for sitting and cell phones for calling, this is for most of us obvious. But even for the one who never seen a chair or a pair of scissors before could probably figure out what it is meant for, in other words the pair of scissors affords the user to cut with it (Norman, 1988). Affordance supports our conceptual modal of what a system or device is able to accomplish. There for, designing for limited affordances, and only affordances that support the intended goal for the device, should support usable devices. Affordance together with constraints when designing things is powerful tools to support usability (Norman, 1988). Though in different situations different devices and systems get different affordances no matter how well designed they are. Due to different context a pair of scissors can become a knife and a chair become table. The device drifts away from the visible affordances towards the invisible and towards new goals.

The drift from the obvious and intended goal to something unforeseen is what Braa et al., (2000) defines as drifting. Here, tactics play a major role as tactics in difference to strategy is dynamic and supports the seize of sudden opportunities that align with our goals.

> "Matching visible and invisible affordances with tactics leads to new uses; re-invention of artifacts and technologies and their shift away from the pre-assigned uses. The result is drifting."

Ciborra, s187, in Braa et al, 2000

Tactics is here seen as the practical part of the new action that together with newly discovered affordances for Invention and results in drifting. These inventions happen in a local context and would hardly exist outside the context where they appear (Braa et al., 2000).

5 Empirical Study

In total we had about twenty different methods to use and from these we decided to test a couple of methods. It was important for us that the methods elected were taking consideration to the environment, rigour/relevance and that it was possible to test them with the same type of tool. We sat down and discussed to what extent each method was rigorous and relevant. Each method was then placed into a Venn diagram so that we could see how they ranged from rigour to relevance (Figure 5.2). We hoped to find methods in each area of the diagram, which we did. Thereafter we tried to sort out methods that ranged from being carried out in a laboratory environment to a more natural environment. We also started to design a test that would take consideration to the variables that we wanted to look closer into (i.e. context and possibilities/lack of possibilities). The test was designed to be able to perform with our equipment available (i.e. DV-camera, MD-recorder).

Method	Lab	Lab/Natural	Natural	Rigour	Rigour/Relevance	Relevance	Possible to test
4		*			*		*
9		*			*		*
10			*			*	*
11		*			*		*
12	*			*			*

Figure 5.1

All methods in figure 5.1 were possible to perform with the test that we constructed (the ones not listed could not be performed with our tool). From these six methods we sorted out one from each category among rigour and relevance. This leaves us with method number 10, 12 and on of 4, 9 and 11. Method number 9 and 11 was too similar to method number 10 in a certain way so we decided to use method number 4, 10 and 12.
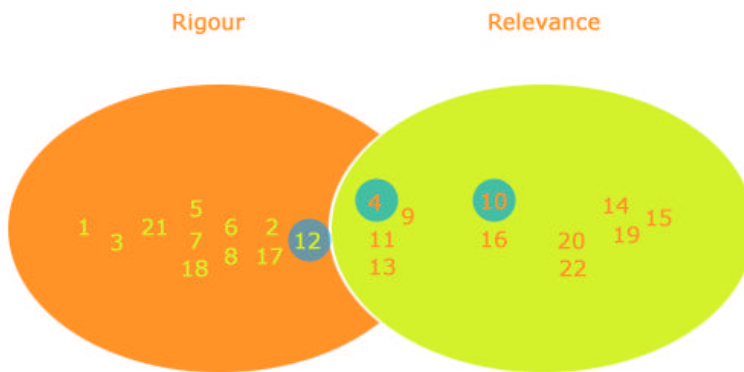
Figure 5.2

We produced three different tasks to evaluate the usability methods selected. These tasks were all designed to be carried out on a PalmV. Each test was quite simple and we estimated that the whole test would be carried out in less than thirty minutes. The tests were not supposed to be used as a test of the PalmV, but rather a tool for us to explore the limits of a static laboratory when it comes to mobile devices. It was also a tool for us to see what information we missed when the mobile device was used in a natural environment. A researcher with usability experience approved the tests that were to be carried out.

The first task was to add a person to the address book. The second task was to schedule two different lessons that were occurring every other week repeatedly for a period of twenty weeks (see Appendix B). The last task was to create a business card. The user supplied their own personal information and transmitted their business card over to another PalmV.

5.1 Performance Measurement

The first method that we evaluated was Performance Measurement (Nielsen, 1993). We engaged five users to participate in our usability tests in the usability laboratory in Aalborg. They ranged from beginners to experienced user and they had very different backgrounds, from a Spanish music composer to an English architect. There were four men and one woman.

The users participated on voluntary basis and they were told that they could interrupt the test at any time if they felt uncomfortable. Before the test took place we introduced them to the laboratory and showed them how the equipment would be used. They were allowed to "play" around with it, all this to make them less nervous and to make them comfortable. We also explained that they were not the subjects of the test, rather we were testing the method.

The laboratory consists of three rooms. One control room where all the technical personal is sitting and controlling the cameras and other effects like background noise and so on (see figure 5.3). One more control room where the test leader is sitting and doing the recording. The test leader is in control of the test situation and helps the user if some problems occur. The control

rooms are placed on each side of a test room. They are separated by windows and were sound isolated.

Figure 5.3

When the user said that they were ready we lead them into the test room (see figure 5.4). Inside there, we told them what they were allowed to do and not. In our case they had to sit in a special angle to the table and they were not allowed to move the PalmV outside specified marks on the table. The three tasks that they were going to do were presented on a laptop in front of them. All usability tests were conducted in one day and recorded on digital video (DV). After the test we asked the each user if we could keep the business card that they had transmitted to us in the last task. We also asked them if we could contact them by mail if we needed to ask the questions that we did not think of during the time we worked with them. All of them were, fortunately, happy to participate.

Figure 5.4

5.2 Co-Discovery Method

The second method to evaluate was Co-Discovery Method (Dumas, Redish, 1993, Rubin, 1994, Lindgaard, 1994). We gathered four new participants. We used the three tasks once again as a tool for evaluating the method. The users sat down at two tables and formed two groups. Each group were given the tasks and told to perform them in pairs on one PalmV. They were told to speak out loud during the test. The tests were recorded on DV (Digital Video).

5.3 Pluralistic Walkthrough

The third and last method that we evaluated was Pluralistic Walkthrough (Bias, 1991). We gathered a new group of PalmV users; in total there were three participants. They ranged from intermediate to advanced users. Once again we used the three tasks as a tool for evaluating the method. We, the authors, acted as moderators and usability experts. Our role was to look at the users while they were performing the tasks and to ask them questions about what they were doing. The users were told to talk out loud and keep up a discussion about what they did and why. After each task we asked them if there was anything to remark upon and if they thought that the task would be able to perform on the run. We also asked them if they would have done it another way if they were on the move. The whole test and discussion was recorded on MD (Mini

Disc).

5.4 Expert Interviews

The expert interviews were all conducted through an e-mail based question form. The questions were more of in the character of "thoughts", and we asked the selected persons to comment on these thoughts. This was done to better explain to them the theme of the thesis. Since all were professionals working in the field of mobility and usability, they all had a deep insight into the theme of this thesis. The "thoughts" we presented to them can be seen as to have influenced them in their answers, therefore compromising the validity and reliability of the interview. Our view is that since this is a group of people with long experience in the theme, they all already have clear view of their field and does not get influenced by our thoughts on the subject.

In September 2000 we sent out e-mails to five researchers within the field of mobility/usability and asked them if they wanted to participate in an interview about Mobile Usability. In early November the four questions were sent out and we asked them to answer before Christmas Eve. We received answers from all the recipients with thoughts and reflections.

The answers was mainly what we had expected and was very much in line with our own thoughts and presumptions, but with some more depth and experience. One of the purposes of asking researcher already in the field was to balance our own lack of experience of fieldwork.
Below are the questions we asked with quotes from some of the answers received.

1.   Mobile usability methods versus conventional usability; is there a need for a whole new method for evaluating mobile gadgets? Is there just a need for an extension of existing methods? Or is there no need at all to make changes to existing usability methods in a mobile setting? [We suggest that there is no need for a whole new methodology concerning MU, but rather an extension to the concept, more like a tactic. We think that you need to consider the ever changing and often-unexpected context and use of mobile gizmos.]

"Yes. Human computer communication with stationary devices is different from human computer communication with "mobile gadgets", hence different methods. The selection and developed of method will depend on what the objective is - so "it depends"."

(Herstad, Jo, 2000)

"I believe that it is more important to establish techniques to capture and evaluate IT use concepts. This is in contrast to the typical CHI community usability study that quantitatively compares the speed of use between two systems. The types of usability study (in a wider sense) that I like is validation in practice."

(Fagrell, Henrik, 2000)

"I think there is a need for extension of existing methods. This problem is more obvious in Mobile usability I think because frequent change of

context etc. but not limited to mobile area only."

(Edvardsson, Senja, 2000)

"I think there is more the world than usability. I think we will see the emergence of use domain specific concepts. E.g. sociability has been proposed and defined (not that great, but its a first shot) as central to the establishment of online communities. This as complementary to usability (which of course still is relevant)."

(Eriksen , Lars Bo, 2000)

"The biggest problem is probably to create a user situation close to reality. Mobile gadgets characteristic are that people use them everywhere. So, the first thing to sort out is how much the context affects the usability of different mobile products?"

(Skov, Mikael, 2000)

2.   Conventional usability methods applied to a mobile setting; what aspects of mobility (if any) does today's usability methods leave out? Also, what are the limitations regarding context, surroundings etc?

"Usually we talk about personal mobility, terminal mobility, session mobility, continuos mobility, discrete mobility and application mobility (from ITU). Depending on what you regard as mobile, the answer will vary :)"

(Herstad, Jo, 2000)

"I do think that traditions usability studies (measure speed of use etc) leave out many important aspects of everyday use of systems."

(Fagrell, Henrik, 2000)

3.   What differences are there between developing completely new gadgets that have no conceptual models and redesigning already existing thingies? In software engineering there is prototyping, would this be a useful method in this case?

"Yes, I believe prototyping may be useful here.…but is there really any new gadgets that have a conceptual model? If you believe that "thingies" evolve, there is really no "new gadgets", but "gadgets" that have evolved from something that already exist - and hence users will have a conceptual model for the use of it."

(Herstad, Jo, 2000)

"Prototyping is definetely also applicable in the development of mobile gadgets, much more than specification approaches. We know so little about how mass audiences perceives mobile gadgets."

(Eriksen , Lars Bo, 2000)

4.   Is there any difference between different types of mobile gadgets when it comes to usability testing, for example between telephones, PDAs and laptops? In other words, do we need separate methods or do you consider it possible to use a "unified" method?

"In FOCUS, Forum for Corporate Usability, at Ericsson - we have a "mantra" which says "it depends". The type of product or solution will affect the method, and the various groups of users will affect the method. In addition, the very obvious fact that it all depends on what the method is to be used for.... There are indeed different methods of testing the same product for teenagers and elderly people for example... "

(Jo, Herstad, 2000)

"The choice of methods is more dependent on what activity/component you want to test on the device then the device itself. So organising tests after device feels wrong."

(Edvardsson, Senja, 2000)

We choose to publish quotes, though some of the quotes are complete answer, to give you as a reader a chance to evaluate the answers for your self. We also wanted to publish the quotes together with questions to give an overview easy to grasp.

5.5 Informal interviews

During the course of writing this thesis we have had many informal interviews with users of mobile devices and experts in the field of usability and mobility. This has been done to see more clearly the problem we are facing creating a new way to test usability of mobile gadgets, but also to learn about the different challenges that a user of a mobile gadget faces in the field.

We would like to point out that we are aware of the limits of our knowledge within the field of usability and especially our practical experience. Also, these small and short tests might be a too small sample to judge upon. Though we find it surprising that we found almost no trace of a context discussion in the books of usability and only brief parts in scientific articles.

6 Analysis

Here we will present our findings from the empirical study of methods. We also present our analysis of the expert interviews and the informal interviews.

6.1 Performance measurement

It became clear to us rather soon that a lab like that was not designed to test mobile gadgets. We had numerous technical problems related to the small size of the gadgets. The cameras used in the laboratory were unable to get a good focus of the gadget. And when we had managed to get an acceptable view of the gadget, we could not move it since it then had been moved out of scope for the camera. We also had problems with the lighting in the laboratory. It constantly gave us reflections in the mobile gadget's display, and thus we could not see what the user was doing with it. This forced us to place the gadget and the person using it in an unnatural way that was nothing like the way they normally would use it.

Another problem not directly related to the technology used was that the test subjects had to read the instructions of what to do in the task. This clearly differs from real world use of a mobile device. You do not always get information that is going to be put into the mobile device in written form.

A third point was that even though we tried to make the subjects feel comfortable and calm, the test subjects did show signs of nervousity, like shaking hands. This of course affected the result of the test.

6.2 Co-Discovery method

This test revealed how a user uses a mobile device in a non-mobile setting, in an office environment. But when the test subjects were asked questions about if they would use the device the same way if they were in another situation, in another context, it became clear that the usage would differ. The test was recorded on DV (Digital Video), but the video was unable to pick up what was going on on the screen, just the conversation and the movement of the test subjects pointing at the screen and discussing elements of the mobile gadget.

6.3 Pluralistic walkthrough

It become clear when doing a pluralistic walkthrough that even the quite experienced users did not know all the "tricks" of the gadget. The test were conducted with people who knew each other well before the test and it became a collaborative learning environment, where the subjects often asked each other questions like "how did you do that", and "I would do that like this".

The time to perform a task varied greatly amongst the users. Also, the subjects learned from each other while performing the tasks. This test was performed indoors in a controlled office environment. The authors often asked the subjects if they would perform the tasks in another way if they had been outdoors, or if they were doing other things at the same time. The answer varied from task to task, but many times the subjects answered that they would do the task completely different "on the run".

This shows that the users use the gadget in different ways depending on the situation. The mobile gadget might work fine in the office environment without stress or other contextual challenging factors, but this does not say much about how it might work in different situations on the run.

We could detect logical faults in the tested applications, and we also found that users can perceive usability matters in completely different ways. A function or feature that one user can not apprehend is completely natural and understandable to others. Users used their gadget in different ways. Everything from starting the gadget to filling in information, the way of doing it differed greatly.

6.4 Expert interviews

The expert interviews clearly confirmed our initial beliefs we had when we began to write this theses; there is indeed a need for research done when it comes to usability in a mobile setting. Also, the traditional usability methods don't take into consideration the context surrounding the usage of the device.

6.5 Informal interviews

The informal interviews were used to broaden our view of the field of mobility, to get the larger picture.

7 Discussion

Below follows a discussion around our findings from our empirical work with usability tests, interviews etc. These findings are in a bigger perspective rather small and only hints to fellow researchers and practitioners within the field of usability. Though we argue that methods developed for certain situations needs to be reconsidered when the conditions changes.

7.1 Thoughts of findings

Like nomads who travel around our community with our gadgets in our breast pocket. From our home to the bus, at work and in the supermarket (Kristoffersen & Ljungberg, 1999). We are indeed mobile - mobile users of mobile technology. Technology design for certain situations and contexts. But with all these different places we go to and daily situations we find ourselves in, are the gadgets really designed for multi-context use, or more important, are they tested for that kind of use?

With the theory of drifting that suggests that artifacts are used in new ways it is almost impossible to test a device in every situation and maybe that is not necessary. Though we argue that totally ignore the psychology of an object (Norman, 1988) that affords multi-context and multi-purpose use is not the way to design usable gadgets. With traditional usability you only test the device in a more or less non-realistic context for tasks you thought it would be used for. In a real world test, with an ethnographical approach you might discover new ways to use a device impossible to discover in a laboratory. This is so far pure speculation and one could argue that traditional usability with methods such as "Talk out loud" would produce reasonable usable products. That is true, but if you are looking for inventing new markets and new services it might not be enough with today's competition within the mobile phone business. As Braa et al., (2000) suggests, innovation is tightly connected to the context and situation where it emerges, we agree and raise the question if a laboratory or group discussion -context is the best environment?

The goal of traditional usability to increase learnability, efficiency, memorability, less errors and satisfaction would still be the same, but needs to be applied to new or modified methods in a mobile situation. Many of the methods mentioned above would be difficult if not to say impossible to use in an open environment but how can we modify these methods to work in a mobile scenario to create the possibility to gather the data we need?

Using a method like pluralistic walkthrough where you ask the users to solve a couple of tests, encouraged to talk to each other and solve the test while the researchers asks questions, revealed in our case, a lot of bad design within the software of the product. These discussions are easy to record, and rigour is relatively high. Rigour is not as high as in the performance measurement method, but on the other side it gives us more relevant result. Even more relevant we have the co-discovery method, but this method is lacking even more rigour. But the problem with mobility is that it is mobile, can be used anywhere, and the point of doing tests on mobile devices in a non-mobile environment is questionable. Our analysis revealed that users used the devices differently

depending on the situation, depending on the context. A method like the co-discovery method can be very relevant testing the gadget for usage in an office environment, sitting down infront of a table. But this only shows us how well the device performs in this type of environment, and nothing about the performance in a more contextual challenging environment. We cannot generalize the result from a test in an office environment and say that it is true for all types of environments and contexts. We need to take the methods out on the field, study real world use, but the methods we tested are hard to apply in a real world situation outdoors. But mobile use makes it hard to record and store conversations. To do that, you need wireless microphones that might feel uncomfortable for the user to wear. You also need video to record how the user handles the device physically and that is not an easy task if you, at the same time, want to capture what happens on the screen. You also do not want to interfere with the user in any way. In doing so you would undoubtedly alter the way the user reacts in a given situation. The user also must feel comfortable with being monitored and recorded to get accurate results from the user.

In this case it is not the methods that needs to be modified but rather our data collection tools that needs to be reconsidered.

When we apply usability methods we try to measure how usable an object is in a given situation. We harvest the data that the method is digging out from the situation and try analyzing it for proper understanding and how to make a more usable product. But is the method bringing the right data to the surface or are we missing something out?

As Fagrell (2000) express it:

> I believe that it is more important to establish techniques to capture and evaluate IT use concepts. This is in contrast to the typical CHI community usability study that quantitatively compares the speed of use between two systems. The types of usability study (in a wider sense) that I like is validation in practice.
>
> [Fagrell, Henrik 2000]

Traditional methods measure speed, error rate and consistency in a product. These are important aspects but they are a very small part of a Human Centered Design. They say nothing about how the product affects the given situation where it interacts or the interaction between humans.

When you leave the controlled environment and move out in the real world to execute the usability tests you meet a more complex world to measure. In the real world there is sound, weather and geography that make it harder to do usability tests and that also affects the use of the gadget you are trying to measure. There are also all these interactions with other humans and gadgets that happen in the real world. All these disturbers make the world more complicated to measure. This is what we are trying to visualize below:
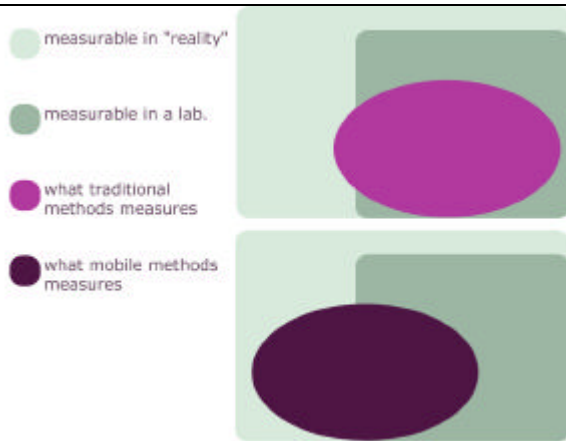
Figure 7.1

This diagram (figure 7.1) visualizes our perception of how context, gadget and usability methods change in two different situations. The traditional usability methods covers a big part of what is possible to measure in a lab but they do not cover much in an unstable environment. The mobile methods cover more in a mobile situation but also a little bit less of what the traditional method does.

So what is it that we miss out in a mobile situation? With the Palm V that we made our tests with it was obvious that the time it took to do a certain task was not paid enough attention. In a real situation when you are writing down a person's address in the Palm while he stands in front of you, seconds feels like minutes.

Also we had trouble with how we would let the users read the task list. The user's concentration was totally focusing on the Palm and on the paper with the tasks during the test. In that situation the task-paper becomes a major actant that do not exist in the real world. In a mobile situation there would be an even greater problem if the user would hold the paper in his hand!

There is of course workarounds to these problems and maybe you only need to be a little creative to solve them. Our suggestion, that we have not tested, is some type of role-play where the user is told to walk down the street and interact with the people contacting him. The people confronting him on his way are of course part of the test. They take on different roles such as an old classmate that the user has not met for a while and the classmate (actor) gives the user his address for him to put into the palm. This creates a much more realistic context for the user, moving away from the unnatural way of reading instructions on a piece of paper or on-screen.

We also see a need for methods inspired of ethnographical methods where we observe the user and the use of a mobile device in a real world situation. This could be done in many several ways. One of the most common would be to let the user observe her self and write it down at a daily basis in a diary. This is one of the methods used in Nielsen and Ramsay's evaluation of WAP in September 2000 (Ramsay, 2000). Taking it a bit further, the next thing to do would be Weilenmann's method of listening to and watching the user when using the mobile device without their knowledge (Weilenmann, Larsson, 2000).

We believe that it is in these types of situations where the device is used in the right context, on the run, while interacting with others and while being carried that you find another set of problems. It also depends on the purpose of use and if the situation for example is under pressure or not.

"Give the palm to, for example, a nurse or doctor at a hospital who were forced to use it as an journal or something, and you will find other faults. If i were to use it right now I do it in one way, pick it up in half an hour and continue. But if the patient could die, it would have another consequence and you would find other types of faults in the gadget."

[Skov, Mikael B., 2000]

Maybe not problems related to efficiency or learnability but more about satisfaction and how it actually feels to use the device. In these situations you might discover that you need to be able to handle the Palm without the stylus because you only got one hand free or that the buttons on the Palms front are pressed down when you carry it in your pocket.

With these solutions for testing in context there is a loss of what we here address as rigour. We loose control over the given situation where the actual test is taking place. The number of factors that possibly affects the test increases and might affect the result in unpredictable way. Though we do not see this as a major drawback. We see control and rigour as a very important factor but not at the price you have to pay when you loose relevance.
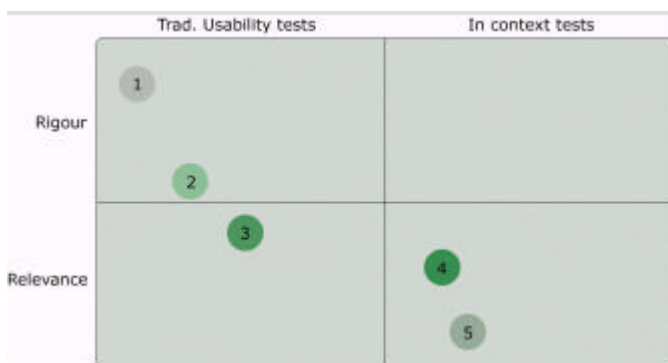

Figure 7.2

1. Laboratory test   **2.** Pluralistic Walkthrough   **3.** Co-Discovery Method
4. Role-Play   **5.** Diary

With the illustration above (figure 7.2) we try shows how different methods are more rigourous and others more relevant. The illustration shows how we would place the different methods. Method number 4 and 5 are the methods we suggest as a better way to also capture context in a usability evaluation.

Most of what is mentioned above might sound obvious for the experienced usability professional. With only a little creativity you would probably think of alternative methods when doing usability tests on mobile devices. But if you study existing literature you will find very little of this creativity in the usability books. There are examples in articles, like Weilenmann, Larsson

(2000) and Ramsay (2000) but nothing gathered in book form that we could find.

	Importance of context-awareness in mobile usability	Need for adapted methods for mobile usability	Traditional usability methods leave important attributes out
Books, Theories			✔
Empiric study	✔	✔	✔
Interviews	✔	✔	✔

Figure 7.3

The above table (figure 7.3) shows our findings from theories, empiric study and interviews and where we found support for our statement. The findings should not be seen as facts but merely a hint towards future research.

8 Conclusion

The first question we had in this thesis was if there was a need for a new or modified method when it comes to mobile usability. As we can see by the answers from out expert interview, there is a great need for doing further research into the field of mobile usability. We conclude that there is no need for developing a whole new method for testing mobile gadgets. Instead we propose a combination of different methods to achieve both relevance and rigour, and to introduce context.

To answer our second questions we propose to introduce methods with a touch of ethnography into the usability testing. Role-playing games where users are in the middle of an act with actors delivering the test. Methods where we observe the anonymous user using the device in an every day situation without any interference what so ever.

The old discussion of rigour vs. relevance continues. We suggest that within mobile usability rigour is important and has a great role when it comes to ensuring consistency between tests and user selections. Though we do find relevance more important in the actual test, which means that rigour is very important before and after the test but during it has to fall away for more relevance.

We are aware that our findings may not be of significance to usability experts but we want to point out that we have been unable to find a discussion about creativity or mobility, context sensitive situations in traditional usability books. Though we found examples of creative usability in scientific articles.

Rigour - Performance measurement in a lab, Relevance - Role playing, ethnographical field studies, contextual inquires. Since our study showed the varying usage of mobile devices among even experienced users, there has to be a strong focus of attention towards testing it in the field with many test subjects.

9 Further Research

This study makes a very good ground for further research within the usability field. Mobile devices will be even more common in the near future and we see a great need for a different design. We will, in the next step of our journey, evaluate our methods of practice mentioned above and compare the result with traditional methods. When that stage, the second, is finished there should be enough empiric knowledge to start creating a framework for design of mobile devices. In future work we will also concentrate of the combined techniques of a PDA and a mobile cell phone. The framework we will try to develop is targeted towards this hybrid of a communication device and a digital filofax.

With methods such as technomethodology, developed especially for the design of artifacts and generalization of human behavior, we will try to define the framework for the design of this hybrid personal mobile device.

In the near future we plan to re-do above mentioned tests in a bigger scale, with at least 20 participants. This time we will be focusing both on the method and the result of the test. This is done to get additional data about the methods and also to harvest data about the actual use of the artifact. This time also mobile phones will be evaluated with the traditional methods.

In parallel with the new tests of PDAs and mobile phones with traditional methods there will be additional tests with above proposed methods such a as role-play, diaries and direct observation.

Role-play is a method sometimes used when designing new artifacts were the test subjects do not have a mental model of such a "non existing" device. The devices we plan to evaluate are existing and we do not use this method because of a weak mental model but rather because of the traditional methods lack of context awareness. A role-play could look like this:

> *"We are standing in front of the local shopping mall. The test subject is told that she will walk through the mall and interact with the persons that confront her.*
> *As she walk through the crowed equipped with a Palm 5 a person approaches and says: - Hello, is that really you??? Linda??? Oh, I haven't seen you since 5th grade, but I have to catch a bus, beam me your address and give you a call…*
> *Here she hopefully picks up the Palm and beam the address over"*

During this conversation someone is recording the interaction on video for later analyses. From this we expect to gain knowledge of how persons handles the Palm under stress and in a quite real situation where we still have the possibility to record the event. We are still in the development of this test and it might be re-designed at a later state. Does it work? Those who live will find out!

Diaries will be used because wants the user to reflect over their use of the device and compare this to how they actually use it in role-plays and in direct observation. The user will write in this diary for two weeks where we also will provide a cell phone or a PDA. If the user is not used to handling such a device we will give a short introduction of critical functions. This because we do not want them to stop using the device because of poor usability. In this case we are not primarily interested in how to make the actual device a more usable product but rather how to make such device truly mobile. To direct the users comments in the direction of mobility we will provide some short questions to consider when writing.

The direct-observation method is quite simple in theory, but intrusive and the ethical aspect can be discussed. When we say direct observation we mean observing the user without the users knowledge, for example, at a café, on the bus or at a shopping mall. Then we record this with either video or just simple notes. From this we hope to gain real use that we can compare with the data from the other methods.

Problems we will encounter are in many ways related to selection of everything from mobile devices to users. The devices we choose to use will have a great affect on the users actual use! For example, a 3Com PalmV affords a different use than an Compaq IPaq PocketPC PDA and the use of a Motorola cell phone will differ from the use of an Ericsson.

When doing traditional tests we have the possibility to choose our respondents. This means that we can have a target group of, let say, technique savvy persons between the ages 15-30. In direct observation it is much harder to have this sort of selection because we do not know whom the user is.

To be able to evaluate our results against traditional usability methods we will compare each result from the proposed methods against the traditional. This means that we will set up a matrix like Figure 9.1 to give an overview of how such a system would look like (the figure is just an example and does not necessarily reflect real findings).

Methods	Laboratory (Rigor)	Contextual Environment (relevance)	Discussion
Performance measurement	X		
Co-Discovery method	X		
Ethnography		X	

Fig. 9.1

With this we will try to show if our proposed methods bring context in to the results and how much that affects the results.

10 Acknowledgements

We would like to gratefully thank a number of people who have in different ways helped us along in the process of writing this thesis.

We thank all the people that we have worked with and have devoted time and facilities for us at the University of Aalborg, Denmark, and especially those at Intermedia. We would especially like to thank Dr. Lars Bo Eriksen and Mikael Skov.

We would also like to thank our colleagues and all the people around at the Laboratorium for Interaction Technology at the University of Trollhättan/Uddevalla for the "moral support" and good ideas. Nothing would be the same without you guys! Especially our mentors at "The lab", Ulrika Lundh Snis, Lars Svensson and Christian Östlund. And of course our Master, Dr. Carsten Sørensen.

Other people who have in some way helped us along in the process of writing this thesis (ranging from proofreading friends and girlfriends and test subjects to helpful usability gurus) are Dr. Henrik Fagrell, Kajsa Warg, Karin Jonsson, Maria Magnusson, Dr. Ben Shneiderman, Martin Elvheim, Senja Edvardsson, Dr. Jo Herstad.

11 References

11.1 Books

Beyer, Hugh, and Holtzblatt, Karen, Contextual Design: A Customer-Centered Approach to Systems Designs, 1997, Morgan Kaufman Publishers, ISBN: 1558604111

Bias, Randolph G., "The Pluralistic Usability Walkthrough: Coordinated Empathies," in Nielsen, Jakob, and Mack, R. eds, Usability Inspection Methods, 1994, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

Braa, K, Sørensen, C and Dahlbom, B. The Planet Internet, Challanges Facing Informatics, Studentlitteratur, Lund Sweden, 2000.

Cohen & Manion (1994). Research Methods in Education, fourth edition

Dumas, JS, and Redish, Janice, A Practical Guide to Usability Testing, 1993, Ablex, Norwood, NJ, ISBN 0-89391-991-8

Eriksson, T. Lars & Wiedersheim-Paul, Finn (1997), Att utreda, forska och rapportera, Stockholm : Liber ekonomi

Freedman, Daniel, and Weinberg, Gerald M, 1990, Handbook of Walkthroughs, Inspections, and Technical Reviews : Evaluating Programs, Projects, and Products, Dorset House, ISBN: 0932633196

Galtung, J. Theory and methods of social research, Universitetsforlaget, Oslo, 1967

Gilb, Tom, Graham, Dorothy, and Finzi, Susannah, Software Inspection, 1993, Addison-Wesley Pub Co, ISBN: 0201631814

Greenbaum, Thomas L., The Handbook for Focus Group Research, 1997, Sage Pubns; ISBN: 0761912533

Hammersley, Martyn & Atkinson, Paul 1995: Ethnography. Principles in practice. London: Routledge.

Holme, Idar & Solvang, Bernt (1997), Forskningsmetodik, Lund: Studentlitteratur

Holtzblatt, K., and Jones, S. "Contextual Inquiry: A Participatory Technique for System Design." in Schuler, D., and Namioka, A. (eds.) Participatory Design: Principles and Practice. Lawrence Earlbaum, Hillsdale, NJ. 1993

Johansson-Lindfors, Maj-Britt, (1993), Att utveckla kunskap - Om metodologiska och andra vägval vid samhällsvetenskaplig kunskapsbildning.- Lund: Studentlitteratur.

Järvinen, Pertti, (1999), *On research methods*.- Tampere, Finland: University of Tampere.

Kahn, Michael, and Prail, Amanda, "Formal Usability Inspections," in Nielsen, Jakob, and Mack, R. eds, Usability Inspection Methods, 1994, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

Lindgaard, G., Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems, 1994, Chapman and Hall, London, U.K. ISBN 0-412-46100-5

Mason, R. O. (1988) Experimentation and knowledge – A pragmatic perspective, Knowledge: Creation, Diffusion, Utilization 10, No 1, 3-24

Mason, R. O. (1989): MIS Experiments: A Pragmatic Perspective. In The Information Systems Research Challenge: Experimental Research Methods, ed. Izak Benbasat, vol. 2. Boston Massachusetts: Harvard Business School Research Colloquium Harvard
Business School, pp. 3-2.

Merriam, Sharan B (1994), Fallstudien som forskningsmetod, Lund: Studentlitteratur

Nielsen, Jakob, Usability Engineering, 1993, Academic Press/AP Professional, Cambridge, MA ISBN 0-12-518406-9

Nielsen, Jakob, and Mack, R. eds, Usability Inspection Methods, 1994, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

Norman, Donald A.(1988). The Design of Everyday Things. MIT Press, London, England.

Repstad, P. (1988), Närhet och distans, Lund: Studentlitteratur

Robinson, B, Doing Case Studies in Educational Research and Evaluation in 'The Open University (1992). Research and Evaluation'

Rubin, Jeffrey, Handbook of Usability Testing, 1994, John Wiley and Sons, New York, NY ISBN 0-471-59403-2

Shneiderman, B. (1998) Designing the User Interface, third ed.. Addison-Wesley Longman, Inc, Massachusetts, USA

Templeton, Jane F., The Focus Group : A Strategic Guide to Organizing, Conducting and Analyzing the Focus Group Interview, 1994, Probus Pub Co; ISBN: 1557385300

Wallén, Göran (1996), Vetenskapsteori och forskningsmetodik, Lund: Studentlitteratur

Wharton, C., Rieman, J., Lewis, C., and Polson, P., "The Cognitive Walkthrough Method: A Practictioner's Guide." in Nielsen, Jakob, and Mack, R. eds, Usability Inspection Methods, 1994, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

Wheeler, David A. (Ed.), Software Inspection : An Industry Best Practice, 1996, IEEE Computer Society, ISBN: 0818673400

Wixon, Dennis, et. al., "Inspections and Design Reviews: Framework, History, and Reflection," in Nielsen, Jakob, and Mack, R. eds, Usability Inspection Methods, 1994, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

Wixon, D., and Ramey, Judith (Eds.), 1996, Field Methods Casebook for Software Design, John Wiley & Sons, Inc., New York, NY. ISBN: 0-471-14967-5

11.2 Articles, papers and proceedings

Beyer, Hugh, and Holtzblatt, Karen, "Apprenticing with the Customer: A Collaborative Approach to Requirements Definition," Communications of the ACM, May 1995.

Boreczky, J., Green, P., Wesselman, H., and Beattie, C. (1990). Wizard of Oz - User Interface Prototyper, Ann Arbor, MI: University of Michigan Transportation Research Institute.

Green, P. and Wei-Haas, L. (1985). The Wizard of Oz: A Tool for the Rapid Development of User Interfaces (Technical Report UMTRI-85-27). Ann Arbor, MI: The University of Michigan Transportation Research Institute.

Holtzblatt, Karen, and Beyer, Hugh, "Making Customer-Centered Design Work for Teams," Communications of the ACM, October 1993.

Johnson, P. (1998). "Usability and Mobility; Interactions on the move." First Workshop on Human Computer Interaction with Mobile Devices. GIST Technical Report G98-1. 21-23rd May 1998. Department of Computing Science, University of Glasgow, Scotland.

Kristoffersen, S. and F. Ljungberg. Mobile Use of IT, In the Proceedings of IRIS22, Jyvaskyla, Finland.

Lewis, C. (1982). Using the "thinking aloud" method in cognitive interface design. IBM Research Report RC 9265. IBM Thomas J. Watson Research Center, Yorktown Heights, New York: Author.

Lindberg, Mikael (1999), Mobil IT i Praktiken, C-uppsats, Uddevalla: Högskolan Trollhättan/Uddevalla - Instutitionen för Ekonomi och ADB.

Nielsen, Jakob, Usability Inspection Tutorial, 1995, CHI '95 Proceedings

Olsson, C, (2000). "The usability concept re-considered: A need for new ways of measuring real web use". Proceedings of IRIS 23, Laboratorium for Interaction Technology

Ramsay, M. and Nielsen, J. (2000). "WAP Usability – Déjà Vu: 1994 All Over Again". Nielsen Norman Group. California, USA

Rowley, David E., and Rhoades, David G. "The Cognitive Jogthrough: A Fast-Paced User Interface Evaluation Procedure." CHI `92 Proceedings, (May 3-7, 1992):

Spencer, Rick. " The streamlined cognitive walkthrough method." CHI 2000 Proceedings, (April 1 - 6, 2000):

Weilenmann, Alexandra & Larsson, Catrine (2000) On Doing 'Being Teenager', Proceedings of IRIS 23, Laboratorium for Interaction Technology

11.3 Electronic sources

Customer Research Groups (2000)
Invented by Gene Lynch and Sue Palmiter of Design Technologies.
http://stc.org/pics/usability/topics/focusgroups.html

Lwin, CPD (1997) Lwin, Thein Designing A Research Study (1997)
http://www.students.ncl.ac.uk/thein.lwin/edd1.html

Nielsen, Jakob, "Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier, 1994" online Web page at
http://www.useit.com/papers/guerrilla_hci.html.

Nielsen, Jakob, "The Use and Misuse of Focus Groups" 1997
http://www.useit.com/papers/focusgroups.html

Appendix

Appendix A

The purpose of this usability test is to evaluate the method of the test. More precisely the goal is to determine how well the method works when testing devices used in mobile settings, if it captures context specific problems.

Problem Statement

3 com Palm	Problem Statement
General	1. Are the external buttons on the front of the palm used to access the device. Why not? 2. Is the user able to start the device without any problems? If there are problems, why? 3. Is the user able to directly grasp the user interface and start working on the task? 4. Is the user able to move freely between applications?
Application, Schedule	1. Does the user create a new event with the "new" button or does she start on the dotted line? Why? 2. Does the user need to consult the help section? 3. Is the user able to open the help menu, under details, without problems?
Application, Address book	1. Is the user able to enter a new person? 2. Is the user able to assign a businesscard 3. Is the user able to beam business card 4. Is the user able to open the help menu, under details, without problems?

Users who take this test should be familiar with the device and platform the test is conducted on. That means that the user should not be a first time user, he or she should at least have a vague conceptual model of the device, how and what to use it for.

Appendix B

T1: Schedule	T2: Address	T3: Beam
1. Start the PDA and open the application "Date book "	1. Start the PDA and open the application "Address book"	1. Start the pda and make the person you just entered in the addressbook your businesscard
2. Add an appointment Tuesday, November 21^{st,} at 15.00 –17.30 with the text "test"	2. Add a new address: Name: Ib René Title: Correspondent Work: 98 12345 E-mail: **ib_rene_cairo@hotmail.com** City: Aalborg	2. Beam your new businesscard to another user
3. The appointment should be repeated every Tuesday for three weeks.		