Mobile Usability - Rigour meets relevance when usability goes mobile

Tomas Lindroth, Stefan Nilsson & Per-Ola Rasmussen
Laboratorium for Interaction Technology, University of Trollhättan/Uddevalla
{tomas, stefan, p-o}@laboratorium.htu.se

Abstract. When we conduct traditional usability tests on applications using stationary computers the context is controlled and equal to the environment where it is to be used. But for mobile devices, with traditional testing there is a risk for irrelevant results since it fails to take the context of its use into consideration. The purpose of this article is to evaluate usability testing methods 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 such as co-discovery method, performance measurement, pluralistic walkthrough and conducted expert interviews with researchers from the mobile as well as the usability field. Together with literature studies and interviews we analyze and discuss around rigour vs. relevance in laboratory and mobile settings. 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, for example a usability lab. We propose a new tactic for usability test mobile gadgets. This is to combine both rigour and relevance in the testing and introduce contextual aspects.

1 Introduction

The mobile device is seen as remote control for business and pleasure where you can buy, sell, control and supervise any artifact or situation. We got more and more mobile gadgets that we carry around everywhere we go, from PDAs that looks like the gadget used in Douglas Adams' novel "The Hitchhikers Guide to the Galaxy", to mobile phones and pagers. For many of us, the functionality of these gadgets is essential to

make us able to work. We demand that they should work whenever, wherever, whatever. But we constantly find ourselves in situations where the usage of these gadgets is compromised. We have to adjust our situation, our context, to be able to use the gadget, for example we have to change our position or eliminate light and sound surrounding ourselves.

As a crude analogy, we can use the situation of using mobile gadgets in different situations with the design of webpages. 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. You need to cover all types of situations the website might be exposed to see how well it functions in all possible scenarios on all combinations of browsers and operating systems. Failing to meet these demands of a website might become be a costly affair.

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.

In this article we study traditional usability methods to see how they function when applied to mobile devices. We want to explore their limitations and see what they reveal about the actual usage of the device.

2 Method

2.1 Method Triangulation

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, 1997) There are also other strengths of using method triangulation. These are for example generalizability and method independence. (Sawyer, 2000) The three methods we used for the triangulation in this article were literature study, expert interviews and empirical tests of usability methods.

2.2 Formal interviews

The formal interviews that we have performed have been structured and sent out by e-mail. 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.

3 Theories

3.1 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)

3.2 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.

The methods we have used 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 Empirical Study

In total we had about twenty different methods to use and from these we decided to test three. It was important for us that the methods selected were taking consideration to the environment, rigour/relevance and that it was possible to test them with the same type of task. Thereafter we tried to sort out methods that ranged from being carried out in a laboratory environment to a more natural environment.

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. 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. 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.

All of these tests were recorder either on Digital Video or on MiniDisc.

4.1 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. There were four men and one woman.

The laboratory consists of three rooms. Two control rooms, one where all the technical personal is sitting and controlling the cameras and other effects like background noise and another 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 the test room. They are separated by windows and were sound isolated.

When the user said that they were ready we lead them into the test room. Inside, 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.

4.2 Co-Discovery Method

The second method to evaluate was Co-Discovery Method (Dumas & Redish, 1993; Rubin, 1994; Lindgaard, 1994). We gathered four new participants and 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 and to speak out loud during the test.

4.3 Pluralistic Walkthrough

The third and last method 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 and once again we used the three tasks as a tool for evaluating the method. The authors acted as moderators and our role was to study 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 be performed on the run. We also asked them if they would have done it in another way if they were on the move.

4.4 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. Since all were professionals working in the field of mobility and usability, they all had a deep insight into the theme of this article. We received answers from all the recipients with thoughts and reflections.

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?

> "…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)

> "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)

> "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)

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 yourself.

5 Analysis

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

5.1 Performance measurement

It became clear to us rather soon that a lab 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 being nervous like shaking hands. This of course affected the result of the test.

5.2 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, but the video was unable to pick up what was happening on the screen, just the conversation and the movement of the test subjects pointing at the screen and discussing elements of the mobile gadget.

5.3 Pluralistic walkthrough

It become obvious 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 several 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 it to filling in information, the way of doing it differed greatly.

5.4 Expert interviews

The expert interviews clearly confirmed our initial beliefs we had when we began to write this article; 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 Discussion

Below follows a discussion around our findings from our empirical work. Usability engineering as a methodology has proved itself to be very useful. Though we argue that methods developed for certain situations needs to be reconsidered when the conditions changes.

6.1 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 are they even tested for that kind of use?

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.

We observed in the tests and the user confirmed in discussions that they would indeed use the device in different ways in different situations. Factors that might provoke different behaviors, expressed by the respondents, could be weather related like rain and temperature and also interaction situations such as talking on the phone or having an face to face conversation.

Also, different users use the device differently in different situations. This might not be surprising but if one user uses the device in different ways in different situations and the use also differs between users we have a problematic situation to analyze. When we say that they used the device in different ways, we mean both physically holding the

device in different ways and navigation through the applications taking different paths. According to these results we find it very hard to draw conclusions about how to measure usability of handheld devices in any way and also to evaluate the methods used. Because we do not no how users use the device there is no way to say what is provoking one behavior from another, if it is the method, situation or the user historical experiences. We see a great need for behavioral studies of the use of mobile handheld device.

The limited groups of people used in this study are, according to us, to small to say anything specific about the use of handheld mobile devices, if the context is not determined by the functionality of the application/device. Which means that a study studying devices with a set of applications designed for use in multi-contextual situations needs a larger population than for example in web usability studies (Nielsen, 1993).

A method like the co-discovery method can be relevant testing the gadget for usage in an office environment, sitting down in front 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.

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!

Also 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.

We 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 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, 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.

7 Conclusion

Our first conclusion is that 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. We propose to introduce methods with a touch of ethnography into the usability testing like role-playing games where users are in the middle of an act with actors delivering the test.

These are 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.

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.

8 Further Research

This study makes a 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 is finished there should be enough empiric data to start creating a framework for design of mobile devices. The framework we will try to develop is targeted towards this hybrid of a communication device and a digital filofax.

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 exist 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…

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.

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. 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 mal. 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.

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.

9 Acknowledgements

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.

References

Bias, Randolph G. (1991), The Pluralistic Usability Walkthrough: Coordinated Empathies. In Nielsen, Jakob, and Mack, R. eds, (1994) Usability Inspection Methods, John Wiley and Sons, New York, NY. ISBN 0-471-01877-5

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

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.

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

Kristoffersen, S. & Ljungberg, F. (1999) Mobile Use of IT, In proceedings of IRIS22, Jyvaskyla, Finland.

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

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

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

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

Olsson, C. (2000), The usability concept re-considered: A need for new ways of measuring real web use. Proceedings of IRIS23, Laboratorium for Interaction Technology, Uddevalla, Sweden

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

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

Sawyer, Steve (2000), Studying Organizational Computing Infrastructures: Multi-Method Approaches. In proceedings of IFIP TC8 WG8, June 9-11, 2000, Aalborg, Denmark

Weilenmann, Alexandra & Larsson, Catrine (2000), On Doing 'Being Teenager'. Proceedings of IRIS23, Laboratorium for Interaction Technology, Uddevalla, Sweden