The role of usability evaluation and usability testing techniques in the development of a mobile system

Vaida Kadytė
Åbo Akademi University
Turku Centre for Computer Science
Lemminkäisenkatu 14 B, 20520 Turku, Finland
Tel: +358-2-2153335
vkadyte@abo.fi

Franck Tétard
Åbo Akademi University
Institute for Advanced Management Systems Research
Lemminkäisenkatu 14 B, 20520 Turku, Finland
Tel: +358-2-2153350
ftetard@abo.fi

ABSTRACT
One characteristic of mobile application development projects is short time-to-market. Short time-to-market implies that very little time is available to application developers between the conception phase of an application and its actual implementation and launching. In the meanwhile, many activities should be conducted, including user requirements elicitation and analysis, application design, testing and evaluation. Along with these activities, a number of decisions will be made, which will influence the design of the user interface. In this paper, we focus on the use of usability testing techniques, and how these influence the design of the user interface in a mobile application development project. We make an account of a usability test, the techniques used, and the results obtained. The paper elaborates on these results with a discussion on how the use of usability testing techniques has influenced the project further on; this discussion is supported by an interview and comments gathered from a technical leader of the development team.

General Terms
Design, Human Factors.

Keywords
Usability testing methods, user interface design, mobile systems

1. INTRODUCTION
Application development of mobile systems is a growing industry. Application development of mobile systems has its own specificities, in the sense that features such as personalization, localization need to be implemented. This means that customer needs need to be understood, and therefore that a user-centred approach should be adopted in order to successfully fulfil user expectations. Also, one specificity of mobile application development is that there are many types of devices: mobile handsets are continuously coming with new additional features to entice users to upgrade, on the other hand, high-speed data capabilities through next-generation cellular networks (2.5G and 3G) trigger the demand creation of more sophisticated mobile phones. This variety of devices means that it is increasingly difficult to develop applications, which will work optimally on all devices. Moreover, mobile application development projects are characterized by short time-to-market, which means that very little time is available to application designers and developers, as well as project management between the conception phase of an application, its actual implementation and launching. In the meanwhile, the application development process should include activities such as user requirements elicitation and analysis, application design, testing and evaluation, to name a few. Along with these activities, a number of decisions will be made, which will influence the design of the user interface. Such decisions encompass e.g. choice of the optimal device for the application, the type of interaction style, or the type of interface. In this paper, we focus on the use of usability testing techniques, and how these influence the design of the user interface in a mobile application development project that was conducted together with a corporate client. We make an account of a usability test conducted during the project, the techniques used, and the results obtained. The paper elaborates on these results with a discussion on how the use of usability testing techniques has influenced the project further on; this discussion is supported by an interview and comments gathered from the technical leader of the development team of the project under scrutiny.

2. PROJECT DESCRIPTION
This paper is based on a research project which aimed at developing mobile business applications for a fine paper value chain. The fine paper production industry is a very mature one, where a long-term relationship with business customers is of particular importance. Recently it has been forced to focus on high quality and innovative products and ability to provide customer care with the help of new ICT capabilities. The project organization involved the third largest fine paper producing company in Europe and one of its key customers - the largest printer in Finland. The initial project goal was to design a mobile system that would provide access to the information services via mobile devices at the point of need, and consequentially would

also benefit B2B relationships in the fine paper value chain. The feasibility study including customer requirement elicitation was conducted in the form of action research, when a team of four people at a research organization participated in the actions of the target organization as consultant body. From the beginning of the application project in March 2003 until the final product was delivered in June 2004, the potential users from both companies were constantly involved in the process of interface design and had influenced its major features during monthly project meetings and workshops.

Even though both companies were very modern in terms of technology and investments in research and development activities, the application development project was a subject to various resource constrains. First of all, we had to decide on a mobile interface design for corporate users - senior and middle level managers - who were both unfamiliar with the mobile business applications and could not afford to spend much of their time on training and evaluation. What was really clear was that they urgently needed to upgrade their existing mobile phones to be able to do more than manage personal contacts and calendars. The business users were on demand to have wireless access to corporate e-mails and databases as well as the capability of running third party applets and services. Smart phones is a new class of handhelds that combine a mobile phone, MP3 players, camera and colour screens with integrated PDA functionalities (calendar, address book, to-do lists), and even the capabilities of running custom applets and accessing corporate databases and currently appears to be the hottest segment of the handheld market. Today these devices not only meet current business needs, but also provide a wide selection base in terms of design, sophistication of features and price, which also makes the usage and usability of mobile applications more complex by different users. In our research project we had a quite challenging goal - to develop a mobile product navigator for novice business users and decide on what kind of smart phone the system would run. Furthermore, the general condition was, that a new corporate standard - the selected smart phone and mobile application - would be easy to learn and usable. Further in this paper, we focus on the use of usability testing techniques, and elaborate on how these influenced the design of the user interface in a mobile application development project.

3. USABILITY TEST DESCRIPTION

3.1 Test objectives

Our first test-goal was to measure the impact of the two different designs on the prospective user (see Figure 1). In particular, we wanted to measure:

- how easy it is for a novice user to learn to use the system on a mobile device
- how easy and efficient it is to operate
- end-user attitudes towards the system

The second test goal was to identify specific problems that the user encountered with the design proposals and with the two devices. We wanted to measure functionality of the systems and users performances within the two system designs on two devices and we wanted to find specific problems that were associated with the usability of the mobile system. Our testing of the interface designs was not concerned with the separate components of the system but concerned more the combination of the components so that we could evaluate how "user-friendly" or good for the purposes the chosen design and device was. We expected that the results of the usability test would help the design team in making decisions regarding further design of the system, and help us giving recommendations whereas which device and which type of interface should be favoured when the system will be taken into use.

Figure 1. Graphical Screen Interface (Grid view) and Hierarchical Screen Interface (List view)

The following general guidelines were set up for the usability test:

1. General questions:
   - On what design is it easier for the user to find the right information?
   - Which design is more intuitive to use?
   - Which design is easier to learn?
2. General goals:
   - A system that is easy to learn and use
3. Quantitative goals
   - Find best performance on device and design, the fastest alternative/user
   - The user should make no errors
4. General concerns
   - Is the system pleasant and easy to use for the purposes it is intended
   - Is the system logical to the actual end-users (e.g. is the menu structure in accordance with their understanding of the product groups)?

The guidelines mentioned above were operationalised as a number of performance measures and subjective measures. Performance measures were directly linked to the quantitative usability goals as well as to the general concerns that were driving the usability test (Dumas & Redish, page 189). Subjective measures, such as opinions, perceptions and judgments of end-users, were linked to the general concerns of the test; these measures were partly operationalised as scores using the PANAS scale (see 3.2), and as qualitative data collected from post-test interviews and by asking users to think aloud during the test(Dumas & Redish, page 187).

The following performance measures were used:

- Time needed to complete a task.
- Number of errors per task.
- Ease of learning: time difference in task completion between two same tasks.

The following subjective measures were used:

- Ease of learning the system.
- Ease of using the products.
- Attitudes towards the system.

3.2 Usability testing methods

We used several data collection methods during the test. Having several data collection methods would ensure that we get evaluation insights of different kinds, which we would feed further into the design process.

User tests: User tests can be conducted in several ways: in the user's natural environment (e.g. on-site testing) or in a controlled environment (e.g. laboratory testing). Both approaches have advantages and disadvantages: choosing the proper environment is very often a matter of trade-off which must be assessed in respect with the objectives of the intended test. In our case, we chose to conduct a laboratory test for two main reasons: (i) we had easy access to laboratory facilities, and (ii) the intended test users, i.e. managers, were willing to join a test in a laboratory. The advantages of running a test in a laboratory were that we could easily control, record, and measure the interaction of the users with the system. Laboratory test planning and data analysis are time-consuming activities: this can be considered to be the main disadvantage when organizing such a test.

User comments: User comments were collected in two different manners. First, users were urged to think aloud when using the system: the think-aloud method is commonly used, although it can be argued that (i) it may distract the users for the task at hand and that (ii) not all users are eager or able to express their thoughts during the course of a test. Second, in-depth interviews were conducted after the test; the purpose of these interviews was to collect feedback mainly about ease of learning the system, and attitudes towards the system.

PANAS scales: PANAS (Positive and Negative Affect Schedule) mood scales have been developed to measure positive and negative affects of individuals. Positive affect (PA) reflects the positive feeling, the extent to which a person feels alert, enthusiastic and active. A high PA means the person is in the state of high energy, full concentration and pleasurable engagement (Watson *et al.*, 1988). A low PA stands for e.g. lethargy and sadness. The Negative Affect (NA) stands for unpleasurable mood states like anger, disgust, guilt, fear and nervousness. A low NA means the person is in a state of calmness and serenity (Watson *et al.*, 1988). The PANAS scales developed by Watson *et al.* enable us to measure PA and NA as two distinct uncorrelated dimensions of affective structure. The PANAS scale is generally known for its stability and it is a relatively easy and trouble free method, short and quick to administer. Some positive and negative affects have been found to be related e.g. to satisfaction and social activity, self-reported stress and poor coping (Watson *et al.*, 1988). We used the PANAS mood scale in order to determine if one of the designs was likely to cause more positive and/or negative feelings than the others.

3.3 Test user selection

The basis for selecting users was that their profile would be as close as possible to the intended user population. For this reason, we aimed at testing managers of the project company and one of their clients. These managers would be somehow literate in computer use, and be familiar with a basic business phone (however, not necessarily with the most advanced business phones available at the time when the test took place). An important selection factor was that they needed to be familiar with the business and the jargon used in the business under scrutiny; this was important as it would ensure that users do not spend time questioning, for example, the meaning of different product groups during the test.

The actual test users of the mobile system are middle level managers. Three of them work at paper producing company (with activities such us product marketing, sales and customer service) and four of them - at a printing company (business customer of the paper company, whose employees are involved in purchasing the paper, warehousing, production planning and control). What unifies the test users is that they all work within the same value chain of fine paper products; they are aware of the complexity of those products and know each other organizational processes very well. On the other hand, the test users have different job focus and responsibilities: four of them have a more business oriented job role (users 1,3,5,7) and three of them are more characterized as technical people (users 2,4,6).

Their average work experience is quite extensive (over 10 years) and indicates that test users are very familiar with the content available for purchasing and processes associated with it. They can be called experts of the content, which is to be navigated via mobile device. Pre-test questionnaires revealed interesting results about their knowledge: since most of the test users know the product information by heart, they have almost never used the product navigator in an electronic format even though it has been available on the web for the last few years. This sort of design of usability test works quite well for the selected fragment of the value chain, because the paper company and printing house operate in a local Finnish market, which is simple and pure and business between them is based on trust and long term relationship.

4. EMPIRICAL RESULTS

The usability test conducted during the project enabled us to gather a lot of data about the users' interaction with the system. We used Noldus usability testing software for recording and analysing observations. As it is not the scope of this paper to make a detailed account of the usability test results, but rather to reflect on the usefulness of usability testing methods and results for interface design, we will present only general statistics and the main insights of the test.

4.1 Data from Observations

Table 1. Observational Data Based on Quantitative Measures

Interface\Device Combinations & Tasks		Task Duration (seconds)		Number of Errors	
		M	SD	M	SD
Joystick List View (JL)	1	122.66	70.90	3.29	3.99
	2	58.10	31.43	0.29	0.49
Joystick Grid View (JG)	1	109.81	98.08	1.14	2.27
	2	84.04	47.73	0.86	1.57
Stencil List View (SL)	1	74.17	25.99	3.00	2.77
	2	38.63	19.03	1.43	2.57
Stencil Grid View (SG)	1	102.73	55.02	1.57	1.27
	2	53.40	24.30	1.29	1.38

Note. SD = Standard deviation, M = Arithmetic mean (n = 7)

Three main insights can be derived from the data presented in Table 1:

- Learning of the user: There is a learning effect taking place. Users perform better (both in terms of task duration and number of errors) task 2 than task 1, and this independently of the device and type of interface.
- Different performance depending on the device: Users perform better when using a device using a stencil rather than a joystick as interaction mode.
- Different performance based on the type of interface: Users perform better using a list-based interface rather than a grid-based interface.

4.2 Mood States with PANAS Moment Instructions

In general PANAS can reflect a general mood state which is fairly similar among the test participants (see Table 2).

Table 2. Positive and negative mood states based on PANAS

	U1	U2	U3	U4	U5	U6	U7	M
Positive Affect (PA) mood states								
JL	2.5	3.6	2.7	2.5	3.4	2.6	2.9	2.89
JG	2.5	3.5	2.5	2.3	3	2.4	2.9	2.73
SL	2.3	3.5	2.6	2.5	3.2	3	3.1	2.89
SG	2.3	3.6	2.4	2.6	3.3	2.5	3.1	2.83
Negative Affect (NA) mood states								
JL	1.3	1	1	1	1	1.2	1	1.03
JG	1	1	1	1	1.3	1.2	1	1.11
SL	1	1.1	1	1	1	1.2	1	1.04
SG	1	1.2	1	1.1	1	1.1	1	1.06

4.3 In-depth Users' Interviews

In-depth interviews were conducted with each user after the test. The main findings of the post task interviews are summarised in Table 3.

To us observers it first seemed that user 1 had more difficulties with using the joystick phone, but the interview and the analysis show that he found this phone easier to use. He found the icons to be moderately descriptive and was of the opinion that for him the picture memory works, you get used to the pictures easily and then you remember them.

User 2 reported few problems with the touch screen phone concerning sensitivity of the stencil, but it was better than the joystick. The results agree with his comments, since he used less time by using the stencil phone. He said that he did not really think about if the icons were descriptive enough but if the system will work in the same way as a normal system on your PC you can organize your desktop so that you have the information you want under your own icons, which is useful. The results show that this user was faster by using the grid layout.

User 3 made a point about the different appearance of the two phones at a meeting with customers. He said that the user would look more professional with the stencil phone at a meeting, because it would look like he would sit and write something, which looks more professional than scrolling on a phone with a joystick looking like you are playing with your phone. The test user preferred icons because he believed icons will be "the thing of the future".

Table 3. Summary of Post-tasks Interviews

	JL	JG	SL	SG
U1	Difficult to get information overview.	⊕	The stencil phone left him feeling unsure, if it will respond to a pen touch or not.	
U2	It is slower than the stencil phone, the screen is smaller		'Too much of scrolling down to find the information' - he was about to give up.	⊕
U3	'It is slow, the screen size smaller and does not look for professional use'.		List view would be far too long – their product list is growing steadily.	⊕
U4	±	Not "his thing", too small, "unresponsive" and "difficult"	±	Would be worried of losing a pen. Can't navigate with one hand.
U5	±	±	±	±
U6	'Information column gets so long that you cannot really use it on the phone'.	⊕	'A pen was a bit misleading: how hard or soft it needs to be pushed; for a correct vertical scrolling – weather to push or to drag a scroll bar; too tiny and can be lost'. Can't navigate with one hand.	
U7	Phone is reacting slower, looks bulky; information bar is not descriptive enough.		List does not look nice	⊕

Note. ⊕ = User expressed clearly his/her preferred combination of interface & device; ± = User doesn't have a clear preference and doubts for corresponding combinations.

User 4 could not reflect his preference for any phone. He stated in the interview that the joystick was not "his thing", he felt it was too small, "unresponsive" and "difficult", even though he did not have any apparent trouble using it. The touch pen appeared to be quite natural for him to use, but at times he got stuck in a situation where he pressed too softly, then too strongly, and became frustrated and clicked a few more times to no avail. In the interview he stated that he felt also the touch pen to be a bit difficult. He also said he would be worried of losing it. He felt that the joystick phone was slower than the stencil, but still he performed his fastest task with the former. User 5 was concerned with the logic arrangement of information in the system and the icons meant "nothing" to him.

Navigation on the phones was somewhat difficult and not intuitive for user 5. Especially the joystick created difficulties for him as he did not seem to find a comfortable way of operating it and slipped quite many times. He also experienced errors when using the touch pen; especially he would hold the pen for too long

on the surface instead of quickly tapping. This made a new menu pop up, which he managed to handle quite independently though. When using the joystick, the screen backlight turned off several times while he was thinking of his next view, which meant that he had to move the joystick in some direction to put the lights back on. This he felt to be a major irritation and something he would really be very annoyed with in the long run. He did not like the icons very much and felt that pictures represented nothing meaningful for him. Otherwise he said, he would like icons, since he is familiar with that metaphor from the computer world.

According to user 6, the joystick phone feels better in the hand than the touch screen alternative. The former is designed so that it should be he held in ones palm. And in deed user 6 behaved a bit more bravely with a joystick phone and immediately took it into his palm.

User 7 liked the touch screen phone at the first glance and didn't change her mind after the test. It was a larger screen size and elegant navigation with a stencil, which gave such a positive impression on her. She also mentioned that a joystick phone didn't look different from other phones she saw, and the joystick itself reminded her a bit of computer games what she never plays.

5. DISCUSSION AND CONCLUSION

The usability test conducted during the project helped the project team to gather a lot of information about how various design options would influence the performance of intended users of the system. It also helped the project team to collect user comments which could be used for further improvements of the system. In the following, we will aim at summarizing the results obtained, the experience gained from using several usability testing methods, and how the results actually influenced the remaining of the project. The discussion presented here is supported by comments of the technical leader of the project; these comments were collected during an informal interview conducted after the project was completed.

Statistical data collected from the observations show clearly that (i) there is a learning effect taking place (comparisons of task 1 vs. task 2), (ii) a list view is more effective than grid view, and (iii) most of the users perform faster with the stencil based navigation than with joystick based navigation. Concerning the benefits of the usability test, we could say that, although we gathered significant and tangible results, the organization of such a test is time and resource-consuming and similar results can probably be obtained with more informal test methods. The main benefit of a usability test remains in the opportunity it gives to the usability experts, development team including technical leader, and project manager all together to monitor many test factors and review these later after the test is conducted.

PANAS was generally better for monitoring the overall mood states of the users during the test: the test results did not reveal any significant individual differences over the positive or negative affects that the different solutions had on the users. As a tool, PANAS with moment instructions was not very useful, since it didn't bring any concrete and reliable insights. Only in three out of seven user cases, PANAS mood states correspond with the users' statements from the post-task interviews. Without in-depth user interviews, interpretation of PANAS results would be difficult and even dangerous.

Post-task interviews with the test participant resulted in rich and subjective information regarding different type of interfaces and input mechanisms. It revealed that three out of seven users had more positive preferences towards using a stencil phone, due to its relatively larger touch screen, faster reaction and elegant look. Two users expressed their clear preference for the phone based on joystick navigation as this kind of method it seemed to be more reliable and more familiar to them. However, the other two users, after listing all the drawbacks for each phone, could not come to the conclusion which one they would prefer. In terms of navigation designs of a product catalogue on a smart-phone, the majority of test users preferred a grid view. We considered unsatisfactory comments from user 4 and 5 concerning a bad organization of information content for the frequent users and suggested an alternative for them – an advanced search was built into the navigator.

User comments collected during the test did not seem to be very helpful in terms of suggestions for design improvements, as they were mostly related with general questions and concerns regarding the phones' features. Use of the think-aloud method proved to be unsuccessful for two main reasons: (i) subjects were not trained to think aloud, (ii) as experts of the system content, users were somehow familiar with the task at hand, and thus they could not easily verbalize their actions, (iii) all of the test participants were Finnish and by nature inherited many characteristics of a high-context culture.

The results of the test are interesting in the sense that user interviews do not necessarily validate the statistical results obtained: it is clear that users perform better using a given device and a given interface, but, surprisingly, these results are not obvious to the users, as they were unable to agree on the best device and the best interface (most users preferred the interface with which they performed worse). The technical leader confirmed that he was surprised that users preferred the grid view to a list view, which proved to be more effective during the test (Technical leader's comments: *"Yes. I did not expect most of the users to prefer the grid view with colourful icons. Another thing is that even if most of the users were performing better and less confused with the list view, they still liked the icon (grid) view more"*). These results did not influence the actual interface design during the project, but the technical leader reflected on how he would overcome this design problem in future designs (technical leader's comments: *"I still believe that you should try to avoid placing pictures or symbols. I believe it is hard for people to associate a certain symbol with particular thing: they have to learn it first, and with time it may be useful to navigate information in such a way. Personally, I do not like to remember symbols, numbers or pictures. I remember the combination and certain placement of information in a certain way. Therefore next time I would develop a product navigator with a grid view, but instead of picture icons, I would have single colour icons with a text below. I also realized that the user interface with icons is very handy with a stencil - you have a larger area and in most cases you will succeed to touch in the right place"*). The technical leader of the project commented that the test results were somehow expected, but did not think that these would influence further design of the system (technical leader's comments: *"Not really (influence). (I) expected most of the responses and anticipated it during the design phase. Different people like different mobile devices as well as webpage layouts. … I also was*

not surprised that people like the stencil phone more since it is faster to navigate: you see and you touch, no need to go up, down or to the left. But of course in certain situations a joystick phone is very handy – you can navigate the phone only with your thumb").

Our technical leader expressed the need for further testing of the system (technical leader's comments: *"It would be interesting to test the ordering system, which is now integrated within the product navigator. The more complex system you have, the more interesting it is to test. Our product navigator has very little functionalities"*).

To summarize, the test proved to be useful to validate several design choices made by the technical leader of the project. From a pure design perspective, the system under scrutiny proved to enhance user performance, although users' opinions and comments seemed to suggest that a less effective design would be preferred. Also, results collected with PANAS moment instruction unfortunately were not confirmed with the interviews; therefore we did not take into account the PANAS scales in our data analysis and decision making.

6. REFERENCES

[1] Dumas J., and Redish J. *A Practical Guide to Usability Testing*. Intellect Ltd. 1999, p. 189.

[2] Watson D., Clark L., and Tellegen A. Development and Validation of Brief Measures of Positive and Negative Affect: the PANAS Scales. *Journal of Personality and Social Psychology*. 1988, Vol. 54, No. 6, pp. 1063-1070.