It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field

Christian Monrad Nielsen
NNIT A/S
Lottenborgvej 24
DK-2800 Lyngby, Denmark
cmne@nnit.com

Michael Overgaard
KMD A/S
Selma Lagerlöfs Vej 300
DK-9220 Aalborg East, Denmark
michael@netmo.dk

Michael Bach Pedersen
ETI A/S
Bouet Moellevej 3-5
DK-9400 Nørresundby, Denmark
mbpedersen@gmail.com

Jan Stage
Aalborg University
Department of Computer Science
DK-9220 Aalborg East, Denmark
jans@cs.aau.dk

Sigge Stenild
Guppyworks
Odensegade 7
DK-2100 København Ø, Denmark
Zilentninja@gmail.com

ABSTRACT
The distinction between field and laboratory is classical in research methodology. In human-computer interaction, and in usability evaluation in particular, it has been a controversial topic for several years. The advent of mobile devices has revived this topic. Empirical studies that compare evaluations in the two settings are beginning to appear, but they provide very different results. This paper presents results from an experimental comparison of a field-based and a lab-based usability evaluation of a mobile system. The two evaluations were conducted in exactly the same way. The conclusion is that it is definitely worth the hassle to conduct usability evaluations in the field. In the field-based evaluation we identified significantly more usability problems and this setting revealed problems with interaction style and cognitive load that were not identified in the laboratory.

Author Keywords
Usability evaluation, field test, laboratory test, experimental comparison.

ACM Classification Keywords
H.5.2 User interfaces (evaluation/methodology).

INTRODUCTION

Usability evaluation has grown into a well-established discipline. The first approaches to usability evaluation as well as today's mainstream methods, e.g. [29], are inherently based on the use of a dedicated laboratory. For several years, this focus on the laboratory has been countered by others who argue in favor of conducting usability evaluations in the field. The discussion of this distinction between field and laboratory has mostly been a matter of opinions, and it has not been prominent in the literature on experimental comparisons of evaluation methods, e.g. [8, 15]. There are, however, examples of experimental comparisons field and laboratory evaluations, e.g. [10].

The advent of mobile devices and systems has revived the controversies about this distinction. Usability evaluation of mobile systems is still an immature discipline [26]. Therefore, basic questions are being discussed. One such question is: should usability evaluation of a mobile system be conducted in the field or in a usability laboratory?

Some argue that a usability evaluation of a mobile system should always be conducted in the field. It is important that systems for mobile devices are tested in realistic settings, since testing in a conventional usability laboratory is not likely to find all problems that would occur in real mobile usage [13]. It also seems to be an implicit assumption that the usability of a mobile system can only be properly evaluated in the field, e.g. [1, 4]. However, usability evaluation in the field is time consuming, complicates data collection and reduces experimental control [2, 13, 16, 18]. There are, however, practical guidelines for handling these challenges [28].

Others argue that usability evaluations in laboratory settings are not troubled with the problems that arise in field evaluations. In a laboratory, the conditions for the evaluation can be controlled, and it is possible to employ facilities for collection of high-quality data such as video recordings of the display and user interaction [3, 16, 17, 19, 30].

NordiCHI 2006: Changing Roles, 14-18 October 2006, Oslo, Norway

The similarities and differences between field and lab-based usability evaluations of mobile systems are beginning to be explored. Some of the comparisons that have been made have observed that there are different interaction behaviors in the laboratory and in the field settings, and they conclude that it is worthwhile carrying out evaluations in the field, even though it is problematic due to difficulties in capturing screen content and the interaction between the user and the mobile device [2, 27].

More recently, contradictory results have appeared. A controversial paper presented at Mobile HCI 2004 concluded that the added value of conducting usability evaluations in the field is very limited and recreation of central aspects of the use context in a laboratory setting enables the identification of the same usability problems [18]. These results are supported by another comparative study where it was concluded that the same usability problems were found both in the laboratory and in the field [14].

The source of these contrary conclusions is not clear. Some of the experiments employ a low number of test subjects. Yet one of the recent experiments is based on 40 users [14]. In most of the experiments, the data collection techniques have not been the same in the field and laboratory tests. This difference is acknowledged in one of the experiments as they state that the dissimilarity in results between laboratory and field evaluation may be a consequence of the differences in quantitative and qualitative data collection techniques [27]. Another example of a significant difference is that task assignments have been used in one setting but not in the other [18].

This paper presents results from an empirical study that was designed solely to enquire into the differences between field and laboratory usability evaluations of mobile systems. The study involved usability evaluation conducted under similar conditions in both a field and lab setting. In order to provide a solid basis for comparison, data collection in the two settings was made with exactly the same equipment. The following section 2 presents the system that was evaluated in the experiment. Section 3 describes the method for the experiment. This includes a description of the equipment that was used to collect data. Section 4 presents the results from the experiment, where the two evaluations are compared in terms of identified usability problems and measurements of usability in accordance with the ISO 9241-11 [12]. Section 5 discusses the results in a broader context, and section 6 provides the conclusion.

SYSTEM DESCRIPTION

The two usability evaluations were made on a mobile system that is used by skilled workers for registering their use of equipment, materials, mileage and time. The system runs on a regular Sony Ericsson T68i mobile phone, with an AirClic barcode scanner attached and uses GPRS for transmitting data. The system is part of a larger administrative system that was not covered in the evaluations.

The user of the system applies a sheet of paper that contains barcodes for tools, equipment, and materials that are used as well as system commands. When a user needs to register some kind of information, he scans the appropriate barcode, which provides access to menus in the system. Figure 1 shows how to execute a barcode scan with the system. Additional interaction with the system is done through the keyboard of the mobile phone.

Figure 1. Using the barcode scanner.

METHOD

Two user-based usability evaluations were conducted, one in a usability laboratory and one in a field setting. Both evaluations were based on Rubin's [29] guidelines for planning and conducting usability tests.

Experimental Design

The two evaluations involved users that were skilled worker apprentices. These apprentices get part of their training in practice and part of it on a technical high school. The evaluations were conducted while they were at the technical high school.

A teacher at the school described initial task proposals, which were then modified to fit the purposes of the evaluations. The teacher was then again consulted in order to ensure that the tasks covered and resembled a real-life working situation. This resulted in nine specific tasks, which dealt with the following working assignments that should be solved using the system:

1. Create a new case with case number and activities in the system.
2. Bring the proper tools for the assignment.
3. Register the mileage used for getting from work to the place of the assignment.
4. Measure a flagstone for the preparation of a stopcock opening (should not be registered) and register the required materials.
5. Lend tools to a colleague on another assignment.
6. Take a break.

7. Make changes concerning the materials used.
8. Continue work on another assignment, which has not been finished.
9. Change the working hours and finish the current day.

The tasks were identical for the laboratory and field evaluation, except for a single task (the actual measurement of a flagstone in task number 4) where the field evaluation included a physical aspect in order to complete the task.

In addition to the tasks, a pre-test questionnaire was made to gather data of the participant's experience with different types of information technology. As a session follow-up a NASA-TLX test [9] was performed alongside a post-test questionnaire. The purpose of the post-questionnaire was to reveal the participant's subjective opinion about the evaluation, the system, and the usage of it.

Two separate teams with a test monitor and a logger conducted the two evaluations. Each team conducted a pilot-evaluation prior to the respective evaluations.

Participants

The test subjects ranged from 16 to 36 in age, and were all apprentices in the field of earthwork engineering. A total of 14 participants took part in the evaluation, and they were divided into two groups of seven. Each group consisted of four from the basic stage of the apprenticeship and three from later stages. The majority of the participants had daily experience with mobile phones. The participants had no or little experience with barcode scanners.

One day before the laboratory evaluation the participants received two hours of training, where they were introduced to the functionality of the system and got a hands-on experience in using the barcode scanner. The training was done by a person who was not otherwise involved in the experiment.

Figure 2. The mini-camera with the mobile barcode scanner system attached.

Data Collection

When a usability evaluation of a mobile system is conducted in the field, it is very challenging to capturing screen content and user interaction [6, 13, 16]. A mini-camera that can be mounted on the mobile phone has been described in the literature, but it was not wireless [25]. We have developed a similar device that is also wireless. The camera with a microphone is mounted on a flexible wire-arm that bends into different positions. The mobile device is attached to the camera holder with Velcro tape, see Figure 2. The camera transmits a wireless video signal to a recorder. This configuration provides steady pictures that enable detailed analysis of screen content and user interaction.

This device was used as the primary data collection equipment. Additionally, system logs with timestamps were generated, which recorded the commands executed by all of the users. These data were collected in exactly the same way in both usability evaluations.

Test Procedure

Before each test session, the test subject answered the pre-test questionnaire. After this, the test monitor gave an introduction to the evaluation. The test subject then worked through as many of the nine tasks, written on paper slips and handed to him one by one, as possible. During the test session, the test subject was thinking aloud. If the test monitor observed that the test subject was helplessly stuck, the evaluation proceeded to the next task, even though the current task was not completed. The test subject was asked to say when he felt he had completed a task. Each session was limited to 40 minutes. After the session, the test subject did the NASA-TLX test and the post-test questionnaire.

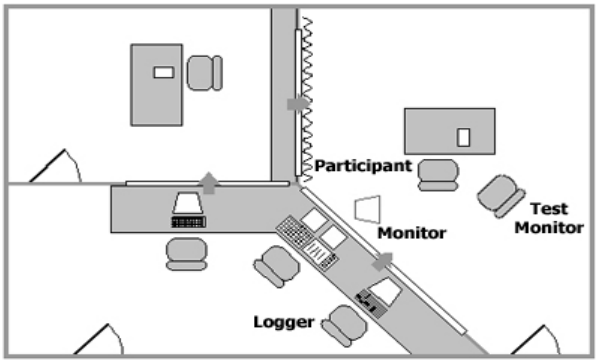

Figure 3. An overview of the usability laboratory.

Laboratory tests

The laboratory tests took place in our usability laboratory. The user was placed at a table and the test monitor was sitting behind him to his right hand side. The logger was placed in an adjacent control room behind a one-way mirror. The participants were given a tour of the laboratory facilities to show them the testing equipment and how the laboratory evaluation would be conducted, so that they would be more familiar with the testing environment.

Four cameras recorded the session; one in front of the test subject and the test monitor, one angled from above, a close-up of the table, and the mounted mini-camera, see

Figure 4. The image from the mini-camera was visible to the test monitor via a monitor placed behind the participant, see Figure 3. A microphone recorded the sound.

Figure 4. The combined camera recordings.

Field tests

The field tests were conducted in a warehouse at the technical high school. The warehouse is designed to accommodate practical learning in the construction business and its interior is similar to real working environments. This made it ideal for our evaluation. The user was placed at a specified working area with the test monitor beside him. During the test, the logger was close by, primarily to observe the evaluation and make notes, and secondarily to operate the recording equipment. No other persons were present in the warehouse during the evaluation. The session was video recorded by means of the mini-camera and a microphone attached on the user. Figure 5 shows a participant during the evaluation.

Figure 5. One of the users solving a task during the field evaluation.

Data Analysis

The two teams completed their evaluations separately. In each team, both members identified and rated the severity of usability problems in order to minimize the evaluator effect [11].

Problem list

Each team divided the test data between its two members, so each of them wrote a session log for half of the test subjects. This was based on the video recordings. Both members then worked separately to analyze each session log and mark places with usability problems. No severity rating was made at this stage. Afterwards the two team members compared session logs and discussed each identified usability problem until consensus was reached. This resulted in a problem list with the identified usability problems and an indication of the sessions in which they occurred. Each problem in each session was then severity rated by the two team members together according to the severity ratings proposed by Molich [20], and the highest rating of an instance of a problem was noted, resulting in a severity rated problem list. This analysis was made separately for each of the two evaluations.

Joint Problem List

In order to compare the evaluations, a joint problem list was made. One member from each team reviewed the two problem lists, and made cross-references between the problems in order to find common and unique problems. These problems were then discussed and elaborated if needed. If a partial overlap between problems was found, the overlap was seen as one problem and the remaining two parts became separate problems. After detailing the problems the severity rating of each problem was reviewed and severity was up- or downgraded if needed. All team members discussed the ratings until consensus was reached. The result was a joint severity rated problem list for both evaluations.

RESULTS

In this section we provide an overview of the problems identified in the two usability evaluations.

Evaluation Type and Number of Problems

The two usability evaluations identified 76 different usability problems altogether. 27 usability problems were categorized as critical, 30 problems as severe, and 19 as cosmetic.

The laboratory evaluation identified 104 occurrences of usability problems and the field evaluation 123 instances. A t-test shows no significant difference, between the two evaluations ($t12=0.83$, $p>0.1$) on this matter. Removing multiple occurrences of the same usability problem, leaves 48 different problems identified in the laboratory evaluation and 60 different problems in the field evaluation, see Figure 6. A two tailed large sample test for population proportions shows that this difference is significant ($z=2.85$, $p=0.006$). Thus the field evaluation identified significantly more usability problems.

A comparison of number of problems, when categorized by severity, shows that the field-based evaluation identified more critical and cosmetic problems.

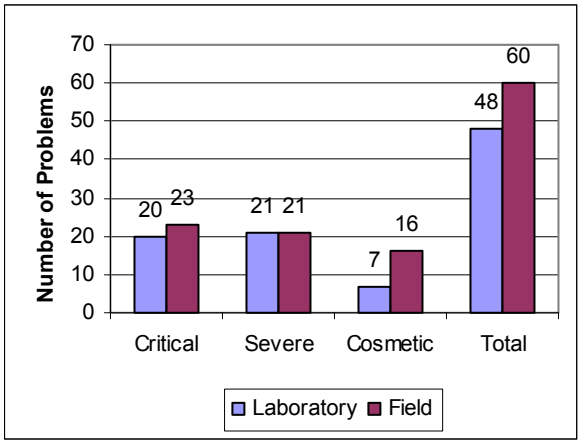

Figure 6. Number of usability problems found in the laboratory and the field evaluation, distributed according to severity categories.

Unique Problems

The evaluations did also uncover problems that only occurred in one of the evaluations. 58% (44 out of 76) of the problems were unique for either the laboratory or the field evaluation and the remaining 42% (32 out of 76) of the usability problems were identified in both evaluations. This result suggests that it might be important to conduct both evaluations, as Pirhonen et al. [27] describe, in order to find the most usability problems. On the other hand the result could indicate that different evaluators identify different problems, as pointed out by Hertzum & Jacobsen [11] and Molich et al. [21].

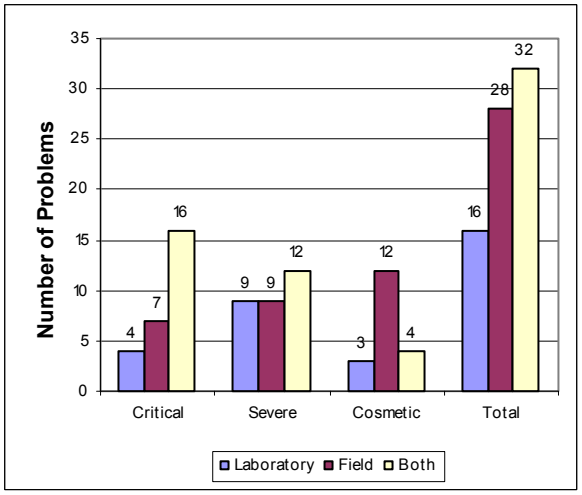

Figure 7. Number of problems, which are unique or found by both evaluations, categorized by severity.

Figure 7 shows that a total of 11 of all the uniquely found usability problems were critical, 18 were severe, and 15 were cosmetic. A two-tailed large sample test for population proportions shows a significant difference in the critical category ($z=1.96$, $p=0.05$) and the severe category ($z=2.24$, $p=0.025$), when comparing number of problems. In the cosmetic category, the difference is very significant ($z=6.19$, $p=0.001$). This indicates that the more severe a problem is, the more likely it is to be identified in both evaluations.

ISO 9241-11

Another way of assessing the usability of the system was by comparing usability according to the ISO 9241-11 standard [12]. The baseline in the tests was that each participant should be able to complete the nine tasks within the 40 minutes time-scope of each session.

Efficiency

The overall completion time for each task was based on the completed instances of a task. Tasks that were not completed were not included.

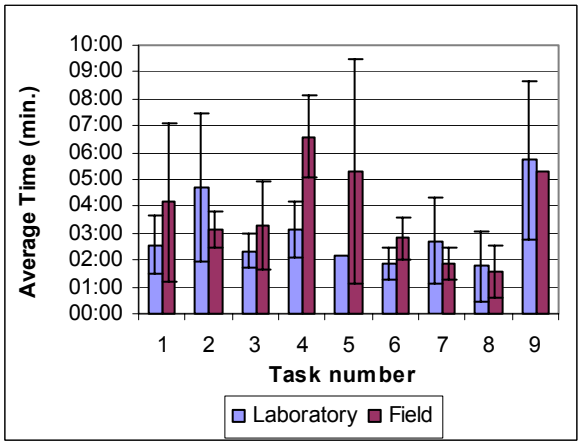

Figure 8. Average task completion time.

Figure 8 shows a comparison between the average time used for each task, in either the laboratory or the field-based evaluation, along with their standard deviations. A t-test shows that the difference in completion time for task 4 was very significant ($t12=4.62$, $p<0.005$). This could be explained by the fact that the participants in the field-based evaluation had an extra aspect to the task, which was measuring of a flag.

Furthermore, there is a significant difference in the time used to complete task 6 ($t12=2.56$, $p=0.025$). Task 5 in the laboratory evaluation was only completed by one participant, which explains the absence of an indication of standard deviation.

Effectiveness

A task was categorized as complete if the end result was equal to a predefined solution. A task was not complete; if the end result differed from the solution, if the task was interrupted by the test monitor, or not started due to the limited time-scope of each session.

A significant difference in number of completed tasks is only present in task 7 ($z=1.67$, $p=0.048$). This indicates that no great distinction exists between the two evaluation

approaches, when looking at the ability to complete the tasks. Figure 9 also illustrates that the least completed task was number 9, which was only completed by 21% (3 out of 14) of the participants. An explanation to this can be the complexity of the task and the time scope of the evaluation. Task 5 in the laboratory was only completed by one participant. This was surprising, as it was one of the simplest tasks, where only one barcode had to be scanned in order to register the lending of a chisel to a colleague on another assignment. The reason is that the majority of the participants did not realize that the same barcode should be used in order to register a tool and deregister the same tool.

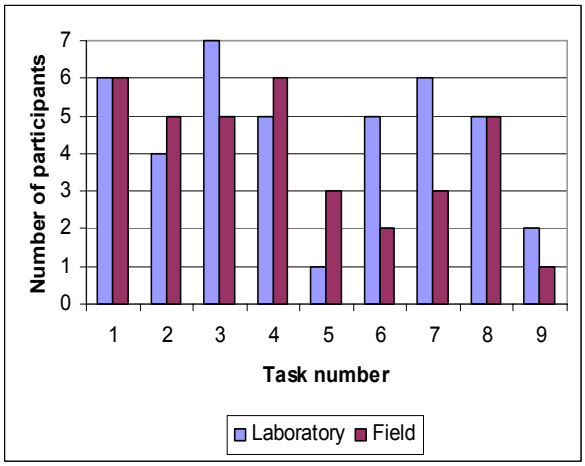

Figure 9. Number of participants that completed each task.

Satisfaction

The participant's satisfaction was measured after the evaluation session by letting them rate their overall satisfaction with the system on a scale from one to seven, where seven was the best. The average in the laboratory evaluation was 5.29 (Std. Dev. 1.28) and 5.00 in the field-based evaluation (Std. Dev. 0.93). The difference is not significant ($t12=0.50$, $p>0.1$). This indicates that the participants' opinion about the system is the same, regardless of the evaluation approach.

Workload

To investigate the workload, as it was perceived by the test subjects in the two settings, a measurement of the workload was made using NASA-TLX scorecards for each participant in the evaluations. The average overall workload for the participants in the laboratory approach was 52.9 out of maximum a score of 100, while the average for the field-based evaluation was 58.4, see Figure 10. A t-test showed that the difference was not significant ($t12=0.63$, $p>0.1$), which indicates that the participants, though being in more realistic settings, did not experience an increased overall workload.

In the NASA-TLX test, the participants also rated how they perceived the mental and physical demands in their test sessions. A t-test reveals a very significant difference ($t12=4.19$, $p<0.005$) in mental demands and a significant difference in frustration level ($t12=2.04$, $p=0.05$) between the laboratory and field-based evaluation, where both aspects were highest in the laboratory. This result is different from the overall workload.

Regarding the physical aspects, there was no significant difference in the way it was perceived by the participants in the two setting ($t12=0.63$, $p>0.1$). It can be argued that the reason for this is that the tasks performed only differ on the physical aspect in task 4. On the other hand, the participants in the field-based evaluation were standing while the participants in the laboratory-based evaluation were sitting, but apparently, that made no significant difference.

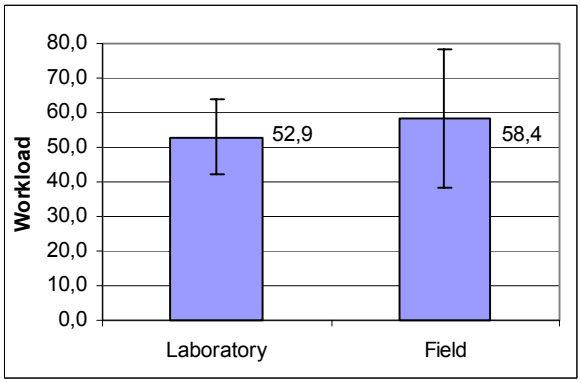

Figure 10. The NASA-TLX result on overall workload.

DISCUSSION

The preceding results provide a quantitative interpretation in terms of the number of usability problems. However, they do not indicate the main kinds of usability problems. In order to facilitate a qualitative interpretation, we categorized the usability problems in terms of a number of themes.

Usability Themes

The themes were identified through a study of relevant HCI literature and were described and acknowledged before the data analysis. Brief definitions of each theme are given below:

Affordance, refers to problems on how the user perceives the properties of an object, and what the actual properties of that object are [24].

Cognitive Load, concerns the amount of cognitive resources needed to use the system [26].

Consistency, relates to consistency in command naming, labels across different screens and consistency in the structure of commands [5].

Ergonomics, relates to the physical characteristics of interaction [5].

Feedback, concerns how the system sends information back to the user about what action has been done [24] and system notifications in relation to system events.

Information, regards how and what information is presented by the system at a certain time [26].

Interaction Styles, covers the design strategy and determines how the system's interactive resources are organized [22].

Mapping, relates to how controls and displays should exploit natural mappings, which take advantage of physical analogies and cultural standards [24].

Navigation, is about how the user navigates through the screens of the system [26].

Task Flow, is about the sequence of steps of which tasks should be conducted [5].

User's Mental Model, The user's model is the mental model developed through interaction with the system [24].

Visibility, concerns which controls are available in the user interface at a specific time [24].

Evaluation Type, Themes, and Severity

The distribution of usability problems on these themes is shown in Figure 11. When looking at the total number of different usability problems identified in the two evaluations, a total of 18 problems were related to *feedback* issues, while 15 problems were related to issues regarding *information*. This means that 43.5% (33 out of 76) of all the usability problems falls within these two themes. A comparison of the laboratory and the field-based evaluation showed that no significant difference exists in the amount of problems, which each type of evaluation identified within these two themes.

The themes *affordance* and *task flow* accounted for 21.0% (16 out of 76) of the problems identified; each theme with 8 occurrences. Looking at the distribution of these problems between the two evaluations, it was clear that problems related to *affordance* were equally present in the both evaluations, while more *task flow* related problems were apparent in the field evaluation. However, the difference between the two evaluations is not significant ($z=1.40$, $p>0.05$).

The remaining problems, 35.5% (27 out of 76), of the total number of usability problems were distributed between the last eight themes. A comparison between the laboratory and the field-based evaluation showed that problems related to *cognitive load* and *interaction style* were identified only in the field evaluation. The reason could be, as described by Baillie [2], that in realistic settings the users more easily become frustrated and thereby the cognitive load is increased. The more realistic use situation in the field could also be the reason for the presence of *Interaction style* related problems in the field evaluation, as the participant had to balance mobile phone and barcodes in his hands, while sometimes having to kneel.

Another comparison between themes and severity categories, revealed that most of the *feedback* and *information* related usability problems were critical. The comparison also revealed that all instances of problems regarding *navigation* were critical, and a significant number of *consistency* related problems were within the same severity category ($z=5.4$, $p<0.001$).

	Laboratory				Field				Distinct Problems
	Critical	Severe	Cosmetic	Total	Critical	Severe	Cosmetic	Total	
Affordance		6		6		4	2	6	8
Cognitive Load						2		2	2
Consistency	3			3	3		1	4	4
Ergonomics	1		1	2	1		3	4	5
Feedback	6	6	1	13	4	7	1	12	18
Information	6	3	2	11	7	4	2	13	15
Interaction Style					1		3	4	4
Mapping		1	1	2			2	2	3
Navigation	2			2	2			2	2
Task Flow			2	2	4	1	2	7	8
User's Mental Model	2	1		3	1	1		2	3
Visibility		4		4		2		2	4
Total	20	21	7	48	23	21	16	60	76

Figure 11. Distribution of problems in relation to usability themes.

Unique Problems, Themes and Severity

A comparison between the unique problems of each evaluation, usability theme and severity category showed that both the laboratory and the field-based evaluation identified several unique *feedback* problems.

Furthermore, the field evaluation identified four unique *interaction style* problems, whereas the laboratory evaluation did not identify a single one. Moreover, the field-based evaluation also identified four unique critical *task flow* problems where the laboratory only discovered one cosmetic *task flow* related problem. The reason for this could be the more realistic context of use, which the field evaluation provided.

Data Collection

Besides presenting insight into the number and nature of usability problems, the evaluations also provided experience in conducting evaluations of mobile systems.

In the laboratory evaluation we combined the mini-camera recordings with recordings from three other cameras. When the recordings were reviewed, it was difficult to see the screen of the mobile phone in detail. It should therefore be

considered which view that contributes most in illustrating an evaluation situation, and make it the main focus on the screen. In the field evaluation, only a full-screen view was available from the mini camera. This provided a good picture of the screen, but made it impossible to properly see interaction with objects in the environment, such as the barcodes. A second camera in the field settings could have provided more information about the user's interaction.

Data Analysis

There were differences between the problem lists from the two usability evaluations. It can be argued that the lists are based on different data, and this may be a reason for the difference in the identified problems. The final result of the two evaluations may also be influenced by the evaluator effect, as analysis of the data requires interpretation by the evaluators. When assembling the joint problem list this effect was noticed. Several problems found in both evaluations were described in different ways or in different detail. By discussing problems, and thereby reaching consensus, this effect was diminished.

ISO Definition

Concerning the ISO usability definition, there are differences in the result of the aspects efficiency and effectiveness, which results in a better overall usability rating of the system in the laboratory evaluation. According to ISO [12], this is not surprising as the context of use influences the usability of a system. This confirms that more realistic context settings in an evaluation provide more valid information about the overall usability of a system.

CONCLUSION

In this paper we have presented and compared the results from two usability evaluations of the same system conducted in two different setting: field and laboratory. By employing identical test procedure and data collection equipment, we have established a solid foundation for comparing these two evaluations.

When the evaluations were conducted in the same way, the field evaluation was more successful as this setting enabled identification of significantly more usability problems compared to the laboratory setting. In addition, it was only in the field evaluation we identified usability problems related to cognitive load and interaction style. This indicates that evaluations conducted in field settings can reveal problems not otherwise identified in laboratory evaluations. Thus the overall conclusion of is that it is worthwhile conducting user-based usability evaluations in the field, even though it is more complex and time-consuming. The added value is a more complete list of usability problems that include issues not detected in the laboratory setting.

The results from the NASA-TLX test show no significant difference between the two usability evaluations in terms of the perceived overall workload. Yet the ratings of the individual factors show that mental demands and frustration level were perceived significantly higher for participants in the laboratory evaluation.

These results are contradictory to recent results on the same issue that are reported in the literature. The reason for this difference may be previous experiments have not used exactly the same experimental procedure and data collection facilities. Our aim was to make the usability evaluations in the two settings as similar as possible. This provided a strong basis for comparison.

It can be argued that the emphasis on similarity reduces the realism of the usability evaluation in the field. The tasks were designed on beforehand, the users were recorded and the recording device made the field evaluation less real. Yet this seems to be a dilemma that is hard to resolve. If we want to compare the two settings, the field evaluation will have to be less realistic.

The findings in this paper are subject to limitations originating mainly from the number of evaluators. In addition, the evaluation focused on novice users of the system. It would be interesting to conduct a similar experiment with expert users. The discussion of the different categories of usability problems relied on a list of themes that were generated from selected literature. It would be interesting to validate these themes through comparison with a broader base of literature.

ACKNOWLEDGMENTS

We thank Net-Mill, Vitus Bering Technical School and Mads Carlsen, Aalborg University for making this experiment possible, and the three anonymous reviewers for constructive and helpful comments.

REFERENCES

1. Abowd, G. and Mynatt, E. (2000) Charting past, present and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction*, 7(1):29-58.
2. Baillie, L. (2003) Future Telecommunication: Exploring actual use, In *Proceedings of IFIP TC13 International Conference on Human-Computer Interaction, (INTERACT '03)*. IOS Press.
3. Bohnenberger, T., Jameson, A., Krüger, A., and Butz, A. (2002) Location-Aware Shopping Assistance: Evaluation of a Decision-Theoretic Approach. In *Proceedings of Mobile HCI 2002*. Springer-Verlag, LNCS.
4. Brewster S. (2002) Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, 6, 188-205
5. Dix, A., Finlay, J., Abowd, G. and Beale, R. (1998) *Human-Computer Interaction*, Prentice Hall Europe, Second Edition.

6. Esbjörnsson M., Juhlin O. and Östergren M. (2003) Motorcyclists Using Hocman Field Trials on Mobile Interaction. *In Proceedings of the 5th International Mobile HCI 2003 conference*. Springer-Verlag, LNCS.
7. Frøkjær, E., Hertzum, M. and Hornbæk, K. (2000) Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*. ACM Press.
8. Gray, W. D. and Salzman, M. C. (1998) Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13(3):203-261.
9. Hart, S. G., and Staveland, L. E. (1988) Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload*. Elsevier Science Publishers.
10. Hertzum, M. (1999) User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests. In *Proceedings of the 5th ERCIM Workshop*, pp. 59-72.
11. Hertzum, M. and Jacobsen, N.E. (2001) The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4).
12. ISO The international Organization for Standardization (1998) *Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability (ISO 9241-11)*.
13. Johnson P. (1998) Usability and Mobility; Interactions on the move. In *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices*. GIST Technical Report G98-1.
14. Kaikkonen, A., Kallio, T., Kekäläinen, A., Kankainen, A. and Cankar, M. (2005) Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, 1(1):4-16.
15. Karat, C., Campbell, R. and Fiegel, T. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems 1992*. ACM.
16. Kjeldskov, J. and Stage, J. (2004) New Techniques for Usability Evaluation of Mobile Systems. *International Journal of Human-Computer Studies*, 60(4-5):599-620.
17. Kjeldskov, J. and Skov, M. B. (2003) Creating a Realistic Laboratory Setting: A Comparative Study of Three Think-Aloud Usability Evaluations of a Mobile System. In *Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction, Interact 2003*. IOS Press.
18. Kjeldskov, J., Skov, M.B., Als, B.S. and Høegh, R.T. (2004) Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings of the 6th International Mobile HCI 2004 conference*. LNCS, Springer-Verlag.
19. Lai J., Cheng K., Green P. and Tsimhoni O. (2001) On the Road and on the Web? Comprehension of synthetic speech while driving. In *Proceedings of CHI'2001*, pp. 206-212. ACM.
20. Molich, R. (2000) *Brugervenlige EDB-Systemer*, 2nd edition. Ingeniøren|Bøger.
21. Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B. (2004) Comparative usability evaluation. *Behaviour & Information Technology*, 23(1).
22. Newman, W.H. and Lamming, M.G. (1995) *Interactive System Design*. Addison-Wesley.
23. Nielsen, C.M., Overgaard, M., Pedersen, M.B. and Stenild, S. (2004) *The Development of a Mobile System for Communicating and Collaborating – An Object-Oriented HCI Approach*, Department of Computer Science, Aalborg University, 2004.
24. Norman, D. (1990). *The Design of Everyday Things*, Doubleday and Company, 2002 Edition.
25. Nyyssönen, Roto and Kaikkonen (2002). Mini-Camera for Usability Tests and Demonstration. Presented in Demo Sessions at the 4th International Symposium on Human Computer Interaction with Mobile Devices, 2002, Nokia Research Center.
26. Pedell, S., Graham C., Kjeldskov J. and Davies, J. (2003) Mobile Evaluation: What the Data and the Metadata Told Us. In *Proceedings of OzCHI 2003*, pp. 96-105.
27. Pirhonen, A., Brewster, S. and Holguin, C. (2002) Gestural an Audio Metaphors as a Means of Control for Mobile Devices. In *Proceedings of CHI'2002*. ACM.
28. Rowley, D. E. (1994) Usability Testing in the Field: Bringing the Laboratory to the User. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press.
29. Rubin, Jeffrey (1994). *Handbook of Usability Testing – how to plan, design, and conduct effective tests*, John Wiley & sons, Inc.
30. Salvucci D. D. (2001) Predicting the Effects of In-Car Interfaces on Driver Behaviour using a Cognitive Architecture. In *Proceedings of CHI'2001*, pp 120-127. ACM.
31. Sannella, M. J. (1994)*Constraint Satisfaction and Debugging for Interactive User Interfaces*. Ph.D. Thesis, University of Washington, Seattle, WA.